Title
The Development of Non-Symbolic Probability Judgments

Author
O'Grady, Shaun

Publication Date
2019

Peer reviewed|Thesis/dissertation

The Development of Non-Symbolic Probability Judgments

By

Shaun O'Grady

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Psychology

in the

Graduation Division

of the

University of California, Berkeley

Committee in charge:

Professor Fei Xu, Co-Chair
Professor Alison Gopnik, Co-Chair
Professor Geoffrey Saxe
Professor Mahesh Srinivasan

Summer 2019

The Development of Non-Symbolic Probability Judgments

Abstract

The Development of Non-Symbolic Probability Judgments

by

Shaun O'Grady

Doctor of Philosophy in Psychology

University of California, Berkeley

Professor Fei Xu, Co-Chair

Professor Alison Gopnik, Co-Chair

Uncertainty plays a role in a variety of early learning processes such as numerical reasoning, language learning, and causal reasoning. Furthermore, adults experience probabilistic data in the course of numerous tasks every day. Thus, the ability to make accurate predictions about future events is a foundational skill which serves both child and adult. Researchers have studied the development of probabilistic reasoning for decades, providing data to suggest that , until early adolescence, children are incapable of accurately predicting future outcomes based on proportion. An equally long scholarly lineage has also provided evidence that adults rely on inaccurate heuristics and biases when reasoning about probability. If prediction is so vitally important to human judgment and decision making, why does the empirical literature suggest humans have impoverished decision making skills when reasoning about uncertainty? What mental representations do humans really on to calculate probability? How do these mental representations change with age and experience?

The current dissertation seeks to answer these questions by studying simple probability judgments made by children and adults. The empirical evidence provided here suggests that children and adults draw on analog magnitude representations of number in order to enumerate sets of outcomes. Furthermore, although both children and adults sometimes use inaccurate heuristics, children rely on these heuristics less with age and adults seem to use them when they perceive two outcomes to be equally likely. In chapter 2, I present findings from a series of experiments employing a non-symbolic ratio magnitude comparison task to investigate the relationship between number approximation, ratio processing, and probability estimation in adults. Empirical results reveal that performance on a probability discrimination task improves as the ratio of the two proportions increases and psychophysical modeling revealed that both numerical and non-numerical stimulus features such as field area, size, and sparsity influence probability estimation. Additionally, these findings reveal that probability estimation is influenced by formally incorrect heuristic decision rules or strategies. Furthermore, findings from two follow-up experiments indicate that these strategies are not influenced by the amount of time participants are given to compute probability and that they persist even when participants are informed that the use of this strategy is not always accurate. While previous research has

investigated the influence of ratio processing and heuristic bias on probabilistic decision making, this series of experiments marks the first attempt to systematically investigate both the psychophysical properties of probability estimation and the factors which influence adults' use of heuristic decision rules in a non-symbolic probability discrimination task. Chapter 3 presents the findings from two experiments designed to investigate the developmental trajectory of children's probability approximation abilities. These results indicate that probability judgments improve with age, become more accurate as the distance between two ratios increases, and that children's perceived probability is influenced by the same psychophysical properties reported for adults (i.e. the size of the objects and the perceived numerosity of target objects). Older children's performance suggested the correct use of proportions for estimating probability; but in some cases, children relied on heuristic shortcuts. Together, these results suggest that children's non-symbolic probability judgments show a clear distance effect, and that the acuity of probability estimations increases with age. In chapter 4, I push this research further by investigating the influence of feedback on children's use of heuristic decision rules. Results from two experiments reveal that children's use of heuristics can be overridden with the proper amount and type of feedback. Together, our findings indicate that children use heuristic decision rules to reason about the outcome of future events but that children can override the use of heuristics if they are provided with enough feedback on trials which conflict with their strategy. These results help shed light on the development of probabilistic reasoning and may lead to improved assessments of children's quantitative reasoning.

Together, the results reported in this dissertation suggest that human probabilistic reasoning is not as impoverished as previous research might suggest. Although children and adults sometimes use inaccurate heuristic decision rules to aid their decision-making, they are also capable of accurately calculating probability based on proportion. Furthermore, children can learn to reformulate their calculations of probability based on feedback and reach a more sophisticated understanding of the proportional nature of probability. These findings have broad implications for a variety of domains such as cognitive development, numerical reasoning, decision-making, strategy selection, and mathematics education.

To my family

Contents

List of Figures

List of Tables

Acknowledgements

I would like to give my sincerest thanks and gratitude to my advisor, Fei Xu, for her guidance and support throughout the course of my dissertation research and graduate studies. As a leader, her mentorship has been an inspiration to me and her compassion and thoughtfulness have been a source of solace and warmth in difficult times. As a researcher, Fei is an exemplary scientist demonstrating an equal measure of curiosity, attention to detail, and creative problem solving. I have learned a great deal about science, cognitive development, life, and parenthood as a result of her dedication and kindness. I would also like to thank my committee members, Allison Gopnik, Geoffrey Saxe, and Mahesh Srinivasan for their insightful comments and supportive guidance throughout the course of this work.

The success of any scientific endeavor is a result of the diligent contributions from a number of collaborators. For this reason, I would like to thank my collaborators, Thomas Griffiths, Ariel Starr, Azzura Rugerri, Zi Sim, Ruthe Foushee, Meg Bishop, Minh-Thy Nguyen, Phyllis Lun, and Harmonie Strohl for their support and insight throughout my graduate studies. I have had the privilege of also working with a number of highly talented undergraduate students at UC Berkeley. The success of this dissertation research is also due to the numerous contributions made by administrators and staff of the Berkeley Psychology Department and the Institute of Human Development including Lisa Branum, Luvy Vanegas-Grimaud, Margaret Bridges, John Schindel, and Tanya Robles.

My gratefulness also extends to my family, notable my dedicated wife Flora and our wonderful daughter Jasmine both of which have been a source of inspiration for my research. I would also like to thank my siblings, Dante Figliolini, Cory Figliolini, and Taylor Dumas whose emotional support is truly impossible to repay. My gratitude also extends to my mother, Ellen Figliolini, and father, John Figliolini, who sadly passed way before the completion of this work. I learned a great deal about hard work and perseverance from both of my parents and I only wish that they could be here to celebrate this work with me.

Finally, I would like to thank the organizations that helped contribute to the success of my graduate research. I am grateful to the financial support provided by the National Science Foundation Graduate Research Fellowship Program as well as the UC Berkeley Regents Fellowship for their financial support over the past five years. I also want to thank the Lawrence Hall of Science, Kensington and Madera Elementary schools, the Bay Area Discovery Museum, and all of the children, families, and adult participants who contributed to this project. Scientific research is one of the few collaborative activities that is as rewarding as it is taxing. This reward comes from sharing discoveries made through the determined efforts of everyone involved. I am grateful to have had the opportunity to share these discoveries with my numerous collaborators and I am excited to learn and grow with all of them.

Chapter 1

Introduction

1.1 Background

Humans experience a great deal of uncertainty throughout their lives. Intuitions about probability and chance are generated from experiences with uncertain events and these notions provide humans with powerful skills for inductive inference. Formal probability is a powerful scientific tool that allows researchers to formally quantify what is unknown or uncertain. How does the untrained human mind respond to uncertainty? How do we cope with the unknown? These broad questions have inspired a research program exploring the quantification of uncertainty via non-symbolic representations of probability and the influence of heuristic decision rules when making decisions based on these quantitative representations.

Probability is formally computed using symbolic notation invented by humans for communicating probabilistic information to other humans. Recent evidence suggests that both humans and non-human animals have surprisingly adept ratio processing abilities (Jacob, Vallentin, & Nieder, 2012; Matthews & Chesney, 2015; Matthews & Lewis, 2017). In the domain of probability, non-human primates can make accurate probability judgments based on proportion (De Petrillo & Rosati, 2019; Rakoczy et al., 2014; Tecwyn, Denison, Messer, & Buchsbaum, 2017) and human infants and children have a remarkably adept understanding of the proportional nature of probability (Denison & Xu, 2014; Falk, Yudilevich-Assouline, & Elstein, 2012; Teglas, Girotto, Gonzalez, & Bonatti, 2007; Xu & Garcia, 2008). Together these findings suggest that there are alternative, non-symbolic methods for quantifying the probability of uncertain outcomes and these non-symbolic alternatives are available to humans before they have the opportunity to learn symbolic representations of probability.

1.2 The mental representation of numerical information

Humans and non-human animals are capable of forming abstract representations of number and researchers often refer to these remarkable numerical abilities as the 'Number Sense' (Dehaene, 2011). One set of numerical abilities relates to our ability to form approximate representations of numerical and magnitude information. These abilities have been demonstrated in both industrialized and non-industrialized cultures (Pica, Lemer, Izard, & Dehaene, 2004) and are thought to be domain general (Dehaene, 2011). One of the most common method for studying the number approximation abilties is through the use of the dot comparison task in which participants view two groups of dots each with a different color and are asked to indicate which group is more numerous.

Using the dot comparison task, researchers have shown that non-symbolic forms of number discrimination follow Weber's law (Halberda & Feigenson, 2008; Whalen, Gallistel, & Gelman, 1999) thus demonstrating ratio dependence: the ability to discriminate two sets of objects based on number depends upon the ratio of the magnitudes of those sets. Additionally, non-numerical features are known to influence ANS representations (Allik, Tuulmets, & Vos,

1991; Durgin, 1995; Gebuis & Reynvoet, 2012; Ginsburg & Goldstein, 1987). Although researchers continue to debate the role of non-numerical features in number approximation (**???**; Leibovich, Katzin, Harel, & Henik, 2017; Odic & Halberda, 2015), recent evidence suggests that human non-symbolic number approximation abilities are best explained by models which integrate both numerical and non-numerical stimulus features (DeWind, Adams, Platt, & Brannon, 2015; Starr, DeWind, & Brannon, 2017). Researchers using this approach systematically vary the size, spacing, and number of dots and then recorded participants' choice strategies. DeWind et al. (2015) demonstrate how a participants' 'bias' for each of these features influenced numerical magnitude judgments. Although researchers studying numerical cognition know a great deal about number approximation abilities, there are surprisingly few studies investigating the link between number approximation and probabilistic reasoning. One of the major contributions of this dissertation is to fill this gap by investigating the mental representation of non-symbolic probability and its relation to number approximation abilities.

1.2.1 Heuristic decision rules in simple probability tasks

Formally, the probability of discrete events is represented as the proportion of target outcomes to all possible outcomes. The discriminability of two non-symbolic ratios is influenced by the distance between those ratios, often referred to as the 'distance effect' (Drucker, Rossa, & Brannon, 2016; Fazio, Bailey, Thompson, & Siegler, 2014; Jacob & Nieder, 2009). A proportion is a special type of rational number in which part whole relations are taken into account during computation. This simplifies the comparisons of probabilities that have different sample spaces. However, when adults are presented with low probability events with equal distributions (i.e. 1 in 10 vs 10 in 100), they typically choose the group with the larger number of target events (Alonso & Fernández-Berrocal, 2003; Denes-Raj & Epstein, 1994; Kirkpatrick & Epstein, 1992; Pacini & Epstein, 1999). Heurisitic based decision making has been a topic of a great deal of research for several decades (Kahneman, 2011; Kahneman & Tversky, 1973; Tversky & Kahneman, 1983) yet very few researchers have studied the task features which influence adults' use of heuristics in simple probability judgments and even fewer studies have investigated the use the developmental origins of heuristics biases.

According to Falk et al. (2012), children use one of four decision making strategies in the 2AFC random draw task. Younger children tend to use 1-dimensional strategies, focusing on a single set of events. They either choose the group with the greater number of target events ('greater win') or they choose the group with the smallest number of non-target events ('lowest loss'). Older children use strategies which can account for both sets of events. These 2-dimensional strategies include choosing the group with the smallest difference between target and non-target events ('greater difference') as well as choosing the group with the highest proportion ('greater proportion'). Data from a sample of 6- to 12-year-old Israeli children suggest that children progress through these strategies as they learn more about the proportional nature of probability and that by age 8 they demonstrate the ability to use the 'greater proportion' strategy. Similar findings have also been reported in a sample of German children using a computer-based design (Ruggeri, Vagharchakian, & Xu, 2018). The second major contribution of this dissertation is to demonstrate that both children and adults sometimes rely on these heuristic decision rules in tasks involving both approximate and exact numerical information.

1.2.3 Development of numerical, proportional, and probabilistic reasoning

This work draws on and contributes to a broad range of research on the development of numerical cognition including non-symbolic magnitude approximation, proportional reasoning, rational number processing, and probabilistic reasoning. I will use these vast literatures to frame the development of several experiments designed to trace the developmental trajectory of non-symbolic probabilistic reasoning. In doing so, I hope to demonstrate the value of the empirical work contained in this dissertation as well as frame future research questions at the intersection of these domains.

Development of number approximation. As mentioned earlier, signatures of the human number sense appear within the first year of life (Dehaene, Dehaene-Lambertz, & Cohen, 1998; V. Izard, Sann, Spelke, & Streri, 2009; Lipton & Spelke, 2003; Xu & Spelke, 2000; Xu, Spelke, & Goddard, 2005). Infant numerical competencies go far beyond large number discrimination. Researchers have shown that infants form expectations about addition and subtraction (McCrink & Wynn, 2004) and are even capable of discriminating ratios (McCrink & Wynn, 2007). Young children are also capable of performing non-symbolic division and multiplication tasks [McCrink and Spelke (2010); McCrink and Spelke (2016)) suggesting ANS representations influence early arithmetical reasoning. Given that these numerical abilities are available at such a young age and continue to develop throughout the life course, it is important to understand how number approximations are used to make sense of the uncertain and probabilistic data that the developing mind must process.

Development of proportional reasoning. For decades, developmental researchers have believed that young children are incapable of accurately reasoning about proportional stimuli (Piaget & Inhelder, 1975; Tourniaire & Pulos, 1985). More recent research has shown that children's proportional reasoning abilities vastly improve over the school-age years (Mix, Levine, & Huttenlocher, 1999; Möhring, Newcombe, & Frick, 2015; Singer-Freeman & Goswami, 2001; Spinillo & Bryant, 1999). Research on proportional reasoning often uses the proportional match to sample task in which participants are first presented with a proportional 'target' stimuli and are then shown several similar stimuli from which they are asked to select the stimulus that matches the proportions of the target. Using this method, researchers have demonstrated that children often make part:part comparisons (i.e. choosing stimilus matches based on matching parts rather than matching the proportion) similar to the heurstic decision rules reported in the probabilistic reasoning literature. Furthermore, recent research investigating proporitonal reasoning with both discrete and continouse stimulus formats has shown that children are capable of making accurate proportional matches when they are presented with stimuli in a continuous format and they tend to show a part:part response bias when presented with stimuli containing discrete, countable parts (Boyer & Levine, 2012, 2015; Boyer, Levine, & Huttenlocher, 2008; Hurst & Cordes, 2018; Jeong, Levine, & Huttenlocher, 2007). These findings fit well with theories of proportional reasoning that claim children's knowledge of the properties of whole numbers interferes with their ability to reason proportionally (Mix et al., 1999; Sophian, 2000; Sophian & Wood, 1997). As a parellel to the developmental literature on number apporximation abilities, these findings also suggest that young children demonstrate accurate proportional reasoning at least when they are presented with stimuli that do not prime them to use their knowledge of whole numbers. Proportional reasoning is also needed when making judgments about simple probabilities invovling binary outcomes and this literature suggests that young children who are capable of sophisticated number approximation and

proportional reasoning should be equally adept at making judgments of probability based on proportional stimuli.

Development of probabilistic reasoning. Classic work investigating school-age children's predictions about single and sequential random draws suggests that children are incapable of correct proportional reasoning in tasks measuring the quantification of probability (Falk, Falk, & Levin, 1980; Piaget & Inhelder, 1975; Siegler, Strauss, & Levin, 1981; Yost, Siegel, & Andrews, 1962), as well as recent research using the methodologically superior 2AFC design (Falk et al., 2012). Young children often choose the group with the greatest number of target objects regardless of the total number of objects in tasks in which they are presented with groups of less than 9 objects and prompted to count (Falk et al., 2012). Furthermore, recent evidence suggests infants (Denison & Xu, 2014) are sensitive to proportions in a 2-alternative forced-choice (2AFC) random draw task. In this task, infant participants watched as an experimenter randomly withdrew a single lollipop from each of two groups of preferred and non-preferred color lollipops (Denison & Xu, 2014). Infants were more likely to approach the lollipop drawn from the distribution with a larger proportion of preferred lollipops even when the total number of lollipops in both groups varied such that the group with the lower proportion actually contained more of the infant's preferred lollipops. Although these findings reveal children's errors in probability tasks involving counting and exact numerical information, very little research has investigated children's reasoning in non-symbolic probability approximation.

1.2.4 Teaching children statistics and probability

As mentioned previously, both children and adults show a bias toward random drawings from groups with a greater number of target events (Falk et al., 2012; Pacini & Epstein, 1999). In the fraction learning literature, this finding is often referred to as the 'whole number bias' (Braithwaite & Siegler, 2017; Ni & Zhou, 2005) and is thought to result from componential processing (Bonato, Fabbri, Umilta, & Zorzi, 2007). When learners compare symbolic fractions, they often simply compare the components of the fraction (i.e. they compare the just the numerators or just the denominators) rather than relating the size of the numerator to the size of the denominator in order to compute the exact numerical value. The integrated theory of mathematical development suggests that these types of errors result from children overextending whole number properties to the set of rational numbers (Siegler, 2016). Based on these findings, one potential explanation for the differences between Israeli (Falk et al., 2012) and German children (Ruggeri et al., 2018) compared to US children is education. In the United States, the Common Core State Standards recommends that educators introduce children to formal probability in the 6th grade (11 to 12 years old) and this introduction usually comes in the form of analyzing outcomes of coin flips, dice rolls, as well as random draw problems.

When children enter the classroom, they have a great deal of prior beliefs and intuitions about a variety of domains and can draw on these beliefs and intuitions during instruction. Contemporary constructivist theories of cognitive development, draw on concepts from Bayesian probability to express developmental change as the integration of prior beliefs with new information (Fedyk & Xu, 2017; Gopnik, 2012; Gopnik & Wellman, 2012; Xu & Kushnir, 2012). Furthermore, education research has shown that learners whose teachers engage with and expand upon children's prior knowledge stand a better chance of understanding part-whole relations in fraction representations of mathematical problems (Saxe, Gearhart, & Seltzer, 1999).

Teachers who understand their students' knowledge state are better able to present information in a manner that will facilitate learning.

What factors influence children's use of heuristics deicision rules and how does teaching formal probability influence intuitive probabilistic judgments? How does a child's prior knowledge of probability interact with the classroom enviornment to influence the learning process? Answering these questions is the third contribution of this dissertation. To my knowledge, only three studies have investigated the effect of feedback and instruction on children's proportional reasoning strategies using the 2AFC random draw task. Fischbein, Pampu, and Mânzat (1970) presented 5- to 13-year-old children with a 2AFC random draw task. On trials containing the same ratio of marbles, younger children systematically chose the distribution with the larger number of target objects. Following instruction, performance on these trials increased to chance levels. Importantly, Fischbein et al. (1970) did not assess children's choice strategies during a pre-test. When performance is at the level of chance, it is difficult to discern whether children learned the correct strategy or whether they were choosing randomly.

A more comprehensive approach was taken by Offenbach, Gruen, and Caskey (1984) who used a computer-based method to present children with probe trials meant to determine which strategy children used on a 2AFC random draw task. Interleaved between these probe trials were feedback trials in which they were shown which response was the correct response. Results revealed that children's use of the proportional strategy improved with age and that both positive and negative feedback did not influence the consistent use of strategies. However, since the feedback trials were interleaved with non-feedback probe trials, children did not receive feedback consistently. Furthermore, since Offenbach et al. (1984) were interested in children's consistent strategy use, they did not vary the feedback trials based on the child's strategy, meaning that all of the children in the study received feedback on the same set of trials regardless of their strategy. One potential explanation for children's consistent strategy use is that there was not enough negative feedback for children using incorrect strategies. If the authors provided more feedback focusing on the specific trials that disconfirm the child's strategy, more children may have learned to change their strategies in response to the feedback.

In a more recent study, Falk et al. (2012) investigated whether children will change their choice in the 2AFC random draw task after viewing the outcome of a random draw. Their results revealed that children's choices were less consistent following a losing draw compared to a winning draw and that this difference declined with age. However, since children were only presented with each trial twice, the authors did not investigate whether this feedback influenced their overall strategy. Furthermore, none of the three studies cited above (Falk et al., 2012; Fischbein et al., 1970; Offenbach et al., 1984) investigated the influence of feedback on specific trial types in relation to the child's prior understanding. Feedback on trials in which the correct choice does not conflict with a child's strategy can inadvertently strengthen the child's belief that they made a correct choice. Furthermore, the uncertainty inherent in probabilistic outcomes can confuse a learner. Thus, the third and final contribution of this dissertation is to demonstrate that children's use of heuristics decision rules is common in both approximation and exact numerical value tasks and that they can learn to override the use of these heuristics if they are provided with the proper amount and type of feedback.

1.3 Précis

Although non-symbolic probability reasoning has been studied for decades, (Chapman, 1975; Falk et al., 1980; Pacini & Epstein, 1999; Piaget & Inhelder, 1975; Siegler et al., 1981; Yost et al., 1962), very little is known about the representational format that humans rely on for computing non-symbolic probability and how these formats are learned and applied throughout ontogeny. How is non-symbolic probability represented in the mind and how do these representations change with age and education? What are the factors that influence the use of heuristic decision rules in probabilistic reasoning tasks and what is the most effective way to teach children about the proportional nature of probability? In this dissertation, I will outline how my collaborators and I have addressed these questions and what implications can be drawn from our results. I see this work making valuable contributions to the literature on numerical cognition and probabilistic reasoning, as well as having broader applications in education and decision making.

Chapter 2 presents a series of experiments conducted with adults investigating the psychophysical properties of non-symbolic probability judgments. The first hypothesis of this work, which I test in a series of four experiments reported in chapter 2, is that people can and do calculate probability based on rapid approximations of the number of possible outcomes in a sample space. Thus, these probability judgments will share signature psychophysical features with the approximate number system, namely, they will show a distance effect (i.e. as two proportions are more distant from each other along the mental number line, they will be easier to discriminate), and they will be influenced by the same numerical and non-numerical stimulus features influencing perceived number such as size and sparsity. Drawing on Signal Detection theory, we compare the predictions of a psychophysical model adapted from the number approximation literature (DeWind et al., 2015) to the response choices of adult participants engaged in a 2AFC random draw task. Results revealed that human adults are capable of rapidly making accurate probability judgments and that these judgments follow Weber's law and are thus characterized by ratio dependence: the closer the proportions were in magnitude, the more difficult they were to discriminate. In accordance with this claim we also find that adults' response data on our task is well fit to a psychophysical model accounting for number, size and sparsity of the visual arrays.

Previous research has shown that adults show a ratio bias effect when reasoning about the outcome of a single random draw when provided with exact numerical value information (Alonso & Fernández-Berrocal, 2003; Denes-Raj & Epstein, 1994; Kirkpatrick & Epstein, 1992; Pacini & Epstein, 1999). Thus, our secondary hypothesis claims that adults sometimes rely on this same formally incorrect decision rule when reasoning about probability approximations that are perceived to be closer in magnitude. Based on this hypothesis, we expect that the accuracy of probability judgments will decrease when the incorrect choice has a larger number of target objects. Importantly, we find a bias in adult judgments toward choices with a higher number of target events and this choice pattern increased with more difficult ratios of proportions. This interaction suggests that people can make correct comparisons based on the exact numerical value of proportions when two proportions are easy to discriminate but they seem to rely on simpler, often incorrect heuristics when two proportions are difficult to discriminate. In follow-up experiments we demonstrate that

Chapter 3 charts the developmental trajectory of non-symbolic probability judgments in elementary school children. In this chapter I will present data from two experiments with 6- to

12-year-old children using similar methods as those presented in Chapter 1. Results suggest that although children are capable of rapidly making accurate probability judgments, their decisions are biased toward choices with a larger number of target events. Importantly, this bias decreases with age until about 12 years in our sample of US children. This finding stands in contrast to that of Falk et al. (2012) who used similar methods to study probabilistic decision-making strategies in Israeli children. Falk et al. (2012) report that around the age of 8, children in their task were capable of correctly using proportion rather than the number of target events to inform their choices. Furthermore, Ruggeri et al. (2018) also found that German children at this age were capable of accurate proportional reasoning on a probability judgment task.

We hypothesize that children in the US use their approximate number system to rapidly compute probability. From this hypothesis, we make three predictions: First, children's probabilistic discrimination abilities will demonstrate ratio dependence (i.e. as two proportions move further apart on the mental number line, they will become easier to discriminate). Second, the ability to discriminate probabilities will improve with age. Third, we predict that probability discrimination will be influenced by the same non-numerical features known to influence ANS representations. Our secondary hypothesis proposes that children's probability choices will be influenced by the same heuristic decision rules reported in previous research (Falk et al., 2012). Based on this hypothesis we predict that children's performance will be influenced by the number of target marbles. Interestingly, data from the current sample suggest that children are capable of reasoning proportionally around the age of 11 to 12 which is the same time that the Common Core State Standards (Best Practices, 2017) recommends US children should be formal introduced to probability in school.

Chapter 4 presents more recent experimental work investigating the factors which influence children's use of heuristic decision rules in simple, non-symbolic probability judgments. Previous research has investigated the role of feedback and instruction on children's random draw choices (Falk et al., 2012; Fischbein et al., 1970; Offenbach et al., 1984) but this previous work did not attempt to track the changes in children's overall pattern of decision making. Using a computerized version of the methods developed by Falk et al. (2012), my collaborators have investigated the role of feedback on children's use of heuristic decision rules.

We hypothesize that children rely on the same heuristic decision rules when asked to make probability approximations as well as when they are asked to make similar judgments based on exact numerical values. Furthermore, our findings indicate that children can learn to abandon these inaccurate heuristic strategies if they are provided with the proper amount and type of feedback with respect to their prior understanding of the proportional nature of probability. Together, these findings suggest both prior knowledge and current information from the learning environment interact to influence children's reliance on inaccurate heuristic decision rules.

Chapter 2

Number Approximation and Heuristics Influence Non-symbolic Probability Estimation

2.1 Introduction

Probabilistic data are ubiquitous in human experience. For example, weather, traffic, and economic information yield a continuous stream of variable data, whether they are reported to us in a news broadcast or gleaned from our own direct experience. What mental processes are involved in calculating probability? Do we rely on numerical information related to rational number when making rapid approximations of probability? In this paper, based on the literatures on probabilistic reasoning and number approximation, we test three hypotheses about the relationship between non-symbolic probability estimation, number approximation, and decision-making: (H1) adults can compute proportions using an analog magnitude system for approximating the number of items in multiple sets; (H2) similar to other types of magnitude approximation abilities, adults' approximate probability estimation will demonstrate characteristics of ratio dependence; and (H3) adults sometimes rely on formally incorrect decision rules (i.e., heuristics or shortcuts) when reasoning about probability. In four experiments, we use a two-alternative forced-choice random draw task in which adult participants are presented with two different groups of red and white marbles. After a brief stimulus presentation time, participants are asked to choose the group which is best for drawing either a red or white marble depending on the participants' assigned condition. Results from this series of experiments have implications for a range of disciplines in cognitive science including numerical cognition, probabilistic reasoning, decision making, visual perception and psychophysics, as well as theories of strategy use in decision making and computational models of strategy selection.

2.1.1 Number approximation and numerical cognition

We hypothesize that humans can approximate the number of target and non-target events in a visual scene depicting multiple sets of events, compute proportions based on these representations, and then use their proportional magnitude judgment abilities to make accurate judgments about probability (H1). Research on numerical processing has shown that both humans and non-human animals are capable of forming abstract representations of number. These remarkable abilities have been termed the 'number sense' (Dehaene, 1997/2011) and

appear within the first year of human life (de Hevia, Izard, Coubart, Spelke, & Streri, 2014; Lipton & Spelke, 2003; Xu & Spelke, 2000; Xu, Spelke, & Goddard, 2005). Analog representations of numerical magnitude consist of a set of modality-independent numerical abilities that allow humans and non-human animals to make rapid approximations of the magnitudes of sets of objects. The rapid and inexact nature of analog magnitude representations are thought to follow Weber's Law (Halberda & Feigenson, 2008; Whalen, Gallistel, & Gelman, 1999) and thus demonstrate ratio dependence: the ability to discriminate two sets of objects based on number depends upon the ratio of the magnitudes of those sets.

Several non-numerical features have been found to influence analog magnitude representations of number. Perceived numerosity is influenced by the distance between the objects (Allik, Tuulmets, & Vos, 1991; Ginsburg & Goldstein, 1987) as well as the size of objects, size of the area occupied by the set of objects (field area), and the density of objects within the field area (Durgin, 1995; Gebuis & Reynvoet, 2012). Although there is still much debate about the role of non-numerical features in number approximation (Leibovich, Katzin, Harel, & Henik, 2017; Odic & Starr, 2018), recent research has integrated numerical and non-numerical features in modeling number approximation abilities of both adults and children to show that numerical decisions are indeed primarily based on numerosity rather that non-numerical features (DeWind, Adams, Platt, & Brannon, 2015; Starr, DeWind, & Brannon, 2017). Importantly, these results suggest that computations involving analog magnitude representations of number should also be influenced by the same non-numerical features.

There is already some evidence to suggest that analog magnitude representations are engaged in numerical computations. Adults and preschool children can perform addition and subtraction based on analog representations of numerical magnitude (Barth, La Mont, Lipton, & Spelke, 2005; Pica, Lemer, Izard, & Dehaene, 2004). Even preverbal infants are capable of tracking the outcomes of approximate addition and subtraction (Chiang & Wynn, 2000; McCrink & Wynn, 2004), which demonstrates that these computations do not require formal schooling or knowledge of numerical symbols . Recent evidence has also demonstrates that both infants and non-human animals can accurately discriminate ratios (Drucker et al., 2016; McCrink & Wynn, 2007). Together these findings suggest that human learners are capable of rapidly enumerating multiple sets of objects (Halberda, Sires, & Feigenson, 2006), discriminating approximate ratios (Drucker et al., 2016; McCrink & Wynn, 2007), and making accurate probability estimation about small, countable, sets of objects (Chapman, 1975). How do humans make accurate probability estimation based on approximate quantities? In order to answer this question, we draw on the literature exploring the psychophysics of number approximation.

2.1.2 Non-symbolic probability judgments

In formal mathematics, the probability of discrete events is computed as a proportion of target outcomes to all possible outcomes. Ratios formalize the relation between any two quantities and proportions are used to test the equality of two ratios. Proportions allow an observer to relate the parts (subsets of outcomes) to the whole (all possible outcomes). Previous research has reported that the discriminability of two non-symbolic ratios (i.e. ratios of objects presented without reference to symbolic representations such as fractions or decimals) is influenced by the distance between those ratios, often referred to as the 'distance effect' and this relationship has been demonstrated in both non-human primates (Drucker, Rossa, & Brannon,

2016) and humans (Alonso-Diaz, Piantadosi, Hayden, & Cantlon, 2018; O'Grady & Xu, 2019; Eckert, Call, Hermes, Herrmann, & Rakoczy, 2018; Fazio, Bailey, Thompson, & Siegler, 2014; Jacob & Nieder, 2009). Based on these findings, we predict that human probability estimation will be less accurate when the probabilities of two possible events are closer in magnitude (H2).

Although probabilistic reasoning has been a productive topic of research for decades, researchers have only recently begun to investigate the psychophysical properties of probability judgments. Fazio et al. (2014) explored the relation between numerical approximation in a dot array comparison task and proportional reasoning in which 5th grade participants (10- to 11-year-olds) were asked to imagine that colored dots on a computer screen were candies and that they were supposed to choose the dot array with the best chances of yielding a target color candy. Although Fazio et al. (2014) were primarily focused on the relations between symbolic number processing, non-symbolic number processing, and mathematics achievement, they did report that number sense acuity in the dot comparison task was correlated with acuity in the proportional reasoning task, which offers some evidence that that the two processes may draw on a similar mechanism. Using a similar design, O'Grady and Xu (2019) were able to chart the trajectory of children's probability approximation from 6 to 12 years of age, and Ruggeri et al. (2018) have shown that the accuracy of children's proportional reasoning is correlated with the acuity of their numerical approximation abilities.

More recently, Eckert et al. (2018) presented chimpanzees and human adults with a 2-alternative forced-choice probability discrimination task in which the subjects were presented with a random draw outcome from two distributions of peanuts and carrots. If chimpanzees can reason accurately about the probability of a random draw, they should choose the outcome drawn from the group with the highest proportion of peanuts, the chimpanzees' typically favored snack. Results revealed a 'distance effect' similar to those found in studies with human children (Fazio et al., 2014; O'Grady & Xu, 2019) and adults (Jacob & Nieder, 2009; O'Grady et al., 2016; Alonso-Diaz et al. 2018): the accuracy of probability estimation for both chimpanzees and humans was influenced by the ratio of the ratios presented. Furthermore, the authors also manipulated the absolute number of peanuts such that on some trials, there were more peanuts in the group with the lower ratio of peanuts in order to investigate the influence of heuristic decision rules such as 'pick the group with the most peanuts' or 'pick the group with the fewest carrots'. While the authors speculated on the role of analog magnitude representations in accurate probability estimation for both chimpanzees and adults, they did not measure the acuity of numerical approximation in their subjects nor did they investigate the influence of non-numerical stimulus features known to influence numerosity judgments.

2.1.3 Use of heuristics in simple probability judgments

Previous work on the ability to calculate probabilities has revealed that participants judgments are biased in predictable ways. In one study, for example, Piaget and Inhelder (1975) devised the random draw task in which participants are presented with two groups of objects asked to choose a group from which to randomly draw a particular object. Using this task, Piaget & Inhelder (1975) found that children used a variety of heuristic decision rules depending on the context of the choice. The most common heuristic reported in this seminal work involved choosing the group with the largest number of target objects even if this group also had the smaller proportion of target objects.

More recent research suggests that children progress from using less complicated, formally incorrect heuristics in which they focus on one dimension of the problem (i.e. 'pick the largest number of target marbles' or 'pick the smallest number of non-target marbles') to using a more complicated yet still formally incorrect two-dimensional heuristics (i.e. 'pick the numerically largest difference between target and non-target marbles'). Findings from this research have shown that by 8-10 years of age, children are capable of accurate, formally correct, proportional reasoning in the two-alternative forced-choice design (Falk et al., 2012; O'Grady & Xu, 2019). Research with adults has also indicated that adult use a similar heuristic decision rule when provided with more complicated decision tasks.

When presented with the same set of proportions based on small, countable quantitates of marbles used to test children, adults typically perform at ceiling (Chapman, 1975). However, when presented with two equal distributions of low probabilities (i.e. 1 in 10 vs 10 in 100) of target versus non-target events, adult participants will often choose the option with the higher number of target events (Kirkpatrick & Epstein, 1992; Pacini & Epstein, 1999), which demonstrates that even adults sometimes revert to single-dimension heuristics. More recent work suggests that adults report feeling more confident about their probability judgments when presented with large compared to small distributions (Alonso-Diaz & Cantlon, 2018). This choice behavior persists even when participants show evidence of knowing that both distributions have the same probability of yielding a target outcome as well as when probabilities are unequal (Alonso & Fernández-Berrocal 2003; Denes-Raj & Epstein 1994). These results indicate that inaccurate heuristic processing pervades human probabilistic decision making.

The current series of experiments seeks to extend the findings from previous reports by including a larger range of probabilities as well as including trials with objects of varying sizes in order to make a more accurate assessment of the role of analog magnitude representations of number in the probability discrimination abilities of human adults. We also seek to investigate factors which influence human adults' reliance on heuristic decision rules. Based on these findings from the empirical literature on probabilistic reasoning (Alonso & Fernández-Berrocal 2003; Alonso-Diaz & Cantlon, 2018; O'Grady et al. 2016), we hypothesize that adults will rely on formally incorrect heuristic decision rules such as 'pick the group with the most [target color] marbles' when their ability to accurately calculate probability is constrained by time and when the proportions are more difficult to discriminate. Thus, we expect that human probability estimation involving a large number of targets and non-targets presented in a short amount of time will be influenced by the number of target objects (H3).

2.1.4 Rationale and synopsis for the current study

In all four of the current experiments, we deployed the same 2-alternative forced-choice in which participants viewed images containing two groups that each contained a mix of red and white marbles. Participants were asked to choose the group that they believed would yield a target color (red or white) from a single random draw. Importantly, in each experiment we created trial types which are meant to distinguish between correct proportional judgments and heuristic biases such as 'pick the group with the greater number of target marbles'. In Experiments 1 & 2 we also deployed a standard dot approximation task in order to measure the acuity of participants' number approximation abilities. In Experiment 2 we altered the size and absolute number of target marbles in order to investigate the influence of heuristic decision rules.

There are two potential explanations for why participants use heuristic biases in this timed approximation task. The first is that the short presentation time may limit their ability to make the appropriate number of approximations. This possibility is tested in Experiment 3 by investigating the influence of presentation time on participants' use of these heuristic decision rules. Another potential explanation for the heuristic bias is that participants did not understand the proportional nature of the task. This was tested in Experiment 4 by including explicit task instructions to the participants meant to inform them that they should not simply choose the option with the greatest absolute number of marbles.

2.2 Experiment 1

In Experiment 1 we presented undergraduate students with images containing two groups of red and white marbles for a very brief period of time. We then asked them to choose the group which they thought was best for randomly drawing a marble of their assigned color. In order to investigate the influence of number approximation ability on probability approximation, we also presented participants with a dot approximation task to assess the acuity of the number approximation abilities.

2.2.1 Methods

Participants. Thirty-seven female and 11 male undergraduate students (N = 48; Mean age = 22.23; SD = 7.64) participated. Sample size was determined based on previous research (Fazio et al., 2014; O'Grady et al., 2016). After providing written informed consent, participants were told that they would be playing two computer-based games related to numerical reasoning.

Material. Images containing two groups of red and white marbles separated by a blue partition were created using Blender 2.72, 3D animation software (http://www.blender.org/). Table 1 provides the proportions of red and white marbles in both groups for each ratio of proportions. Two different types of trials were included. In the *total equal* trials, both groups had an equal number of marbles. These trials can be considered relatively easier because they only require a comparison of the number of marbles of the target color. Images for the *target equal* trials contained groups which had an equal number of target color marbles but different numbers of non-target marbles such that the proportions of marbles in the two bins matched the proportions of the corresponding *total equal* trial. Since the total area of the two groups was smaller for the 'correct' choice, an additional 10 foil trials were created to reduce the chances that participants would learn to simply choose the smaller of the two groups of marbles. Foil trials contained an unequal total number of marbles in both choices like the *target equal* trials with a larger number of target marbles in the 'correct' choice similar to the *total equal* trials.

Table 2.1 Proportion presented in each trial of Experiment 1.

Proportion group 1	Proportion group 2	Ratio of Proportions
0.55	0.50	1.10
0.70	0.60	1.17
0.55	0.45	1.22
0.80	0.60	1.33
0.80	0.55	1.45
0.60	0.40	1.50
0.70	0.40	1.75
0.55	0.30	1.83
0.60	0.30	2.00
0.70	0.30	2.33
0.80	0.30	2.67
0.75	0.25	3.00
0.70	0.20	3.50
0.80	0.20	4.00
0.90	0.15	6.00
0.80	0.10	8.00
0.90	0.10	9.00
0.50	0.05	10.00
0.55	0.05	11.00
0.70	0.05	14.00

Note. Ratios of Proportions are rounded to 2 digits. See the Supplemental Material for a full table including the numbers of marbles used in each group for each trial type.

Procedure. Participants were seated about 60 cm from a MacBook Pro laptop (OSX; Screen resolution 1280 x 800) and were told that they would see some images with two groups of red and white marbles on the screen. Experimenters informed the participants that their task was to collect either red or white marbles depending on the condition to which the participant was assigned. Next, they were told that the computer would randomly select a marble from one of the two groups on the screen and that they could collect marbles by telling the computer which group to draw a marble from. Finally, participants were told that there was always a 'best' choice and that while some of the trials will seem easy other trials may be more difficult and if they were uncertain about which group to choose, they should try to make their best guess. Four practice trials were used to provide an example of how to play the game. For each of these practice trials participants saw two groups of marbles, one contained all red marbles while the other contained all white marbles. Participants were told that the practice trials were intentionally easy and were only meant to familiarize them with the operation of the game.

Images were presented using a script written in MatLab programing language with the psychophysics toolbox (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997). Each participant was presented with 40 test trials and 10 foil trials in one of two semi-randomized orders. Previous research using a similar design presented images for 1320ms for 11-year-olds

(Fazio et al., 2014). Adults have been shown to have faster and more accurate number approximation abilities than children and for this reason we chose to present images in the current study for 750 ms. After stimulus presentation, marbles on the screen were replaced by two bags labeled with a blue '1' or a green '2'. Participants made their choice by pressing one of two keys marked with either a blue or a green sticker corresponding to the bags. After fifteen trials, participants saw a brief animation and were encouraged to take a break from the game. For each trial the computer recorded the participant's choice as well as reaction time. After the game ended and data collection was complete participants saw a screen containing 32 marbles that matched their target color and were told that these were the marbles that they had collected during the game. Figure 1 presents a visual schematic of the procedure for the 2-alternative forced-choice random draw task.

After completing the probability judgment task participants were also asked to perform a number approximation task (Panamath.org; Halberda & Feigenson, 2008) for 10 minutes. Instructions for this portion of the experiment were provided based on the procedure for testing adults prescribed by the Panamath website. In this dot approximation task, participants are presented with an image containing various ratios of blue and yellow dots and are asked to indicate whether there are more blue dots or more yellow dots. Following Halberda, Mazzocco, and Feigenson (2008) a psychophysical model was used to compute each participant's Weber Fraction, an index of numerical acuity. Lower Weber Fractions indicate better acuity, such that participants are capable of making finer discriminations between the numbers of sets of objects. Once the number approximation task was completed, the experimenter debriefed the participant and discussed the rationale for the study.

Statistical analyses were performed using the R programming language (R Core Team, 2008). De-identified data, methods, materials, experimental code, and manuscript preparation code for all four of the experiments in this manuscript can be found on the Open Science Framework at the following link: (https://osf.io/adms6/?view_only=0c28c60a5993466d894273b003b75ab1).

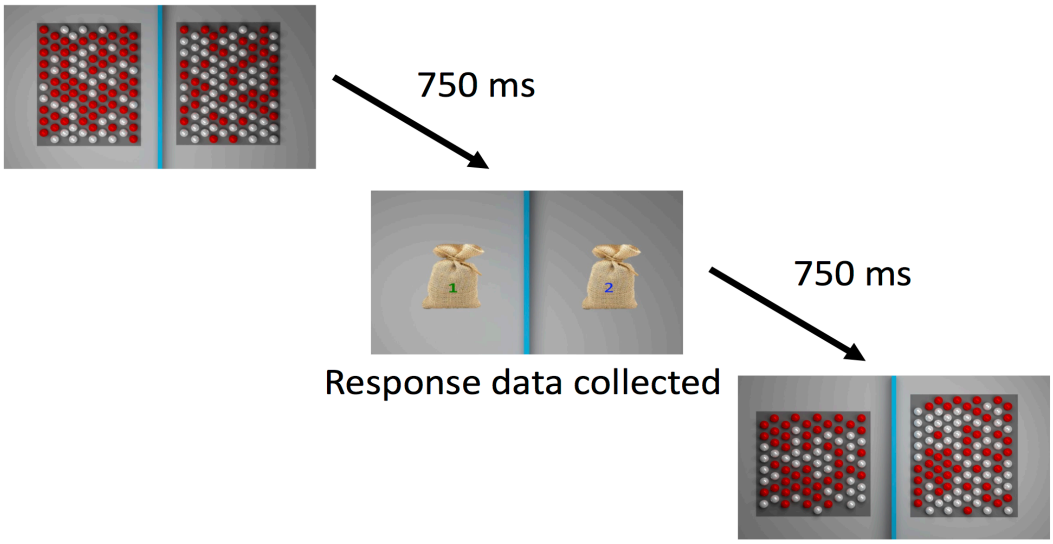

Figure 2.1 Diagram of the experimental procedure used in Experiment 1. The sample image at the top presents a target equal trial while the sample image at the bottom presents a total equal trial.

2.2.2 Results

Analyses of the foil trials revealed that adults were at ceiling on these trial types indicating that they were not simply choosing the group that occupied less space in the image ($M = 0.98$, 95% CI [0.96, 0.99], $t(47) = 65.93$, $p < .001$). Foil trials were not included in subsequent analyses. Additionally, there were no significant differences between participants who were instructed to collect red marbles and those who were asked to collect white marbles nor were there any significant differences between the two counterbalanced orders of trials, so these variables were dropped from subsequent analyses.

General Accuracy. Overall, participants did quite well, performing significantly above chance on both *total equal* ($M = 0.98$, 95% CI [0.96, 0.99], $t(47) = 71.99$, $p < .001$) and *target equal* trials ($M = 0.92$, 95% CI [0.90, 0.95], $t(47) = 32.12$, $p < .001$). Figure 2 displays average performance by ratio of proportions and trial type.

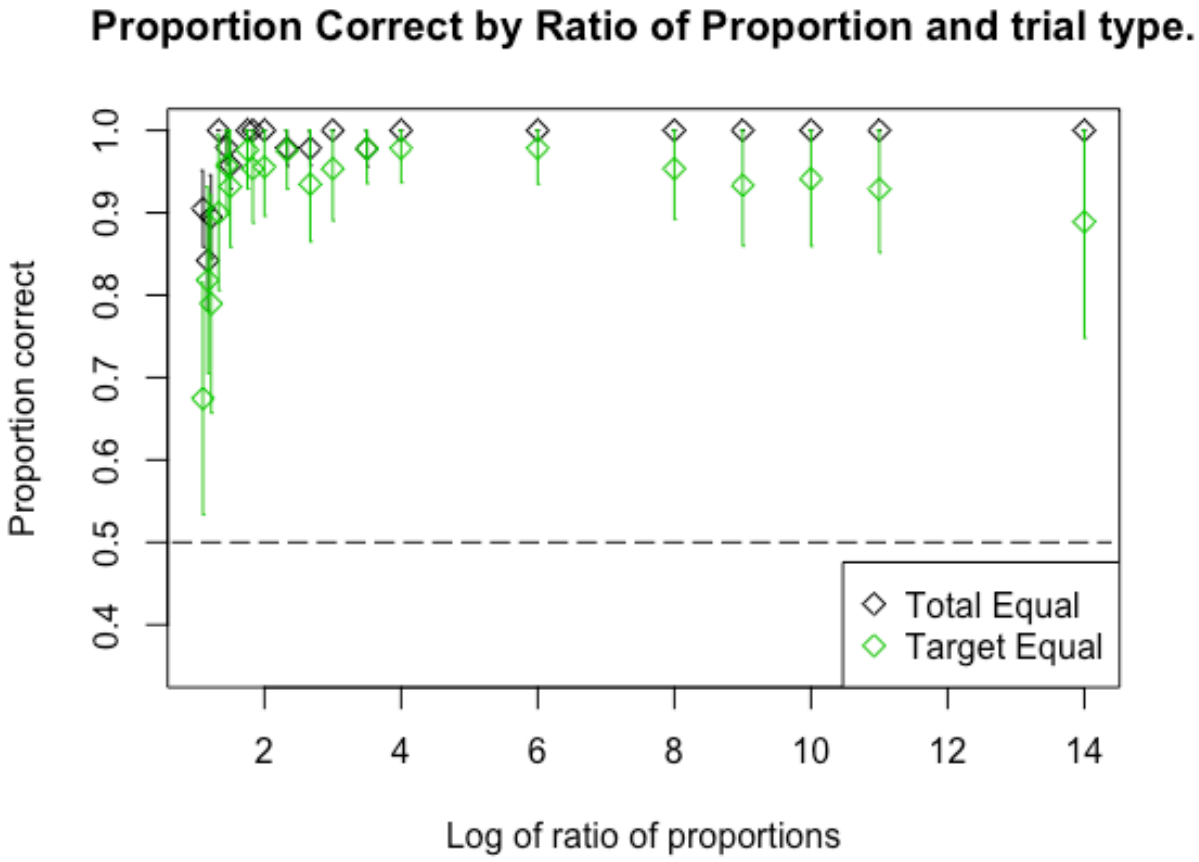

Figure 2.2 Average performance by ratio of proportions and trial type. Error bars indicate bootstrapped 95% confidence intervals.

Statistical Modeling. Generalized Linear Models with Mixed effects (GLMMs) from the lme4 package (Bates, Maechler, Bolker, & Walker, 2015) were used to predict the binary response variable based on experimental, procedural and subject variables. This analysis is the preferred method compared to simple t-tests because regression models can assess the influence of both categorical and continuous variables as well as the interaction between the two on performance. Reaction times were analyzed via comparisons of Linear Models with Mixed effects (LMMs) from the same lme4 package. In both cases, participant identification numbers were entered as random effects in order to account for multiple comparisons resulting from two trial types for each ratio of proportions. Importantly, model comparisons did not reveal any effect of age, gender, order of trials, or the target color.

Comparisons of GLMMs revealed that the model predicting the participant's response from ratio of proportions, trial type, as well as the interaction between ratio of proportions and trial type was found to have the best fit to the data (AIC_{RP*TT} =596.66). This model outperformed the null model (AIC_{null} = 654.65; χ^2= 63.99; *df* = 3; $p < .001$) and the models predicting performance from the ratio of proportions alone (AIC_{RP} = 641.38; χ^2= 48.72; *df* = 2; $p < .001$), trial type alone (AIC_{TT} = 625.56; χ^2= 32.90; *df* = 2; $p < .001$), as well as the ratio of proportions and trial type with no interactions (AIC_{RP+TT} = 613.34; χ^2= 18.68; *df* = 1; $p < .001$).

The interaction term of the model ($\beta_{Interaction}$ = 1.49; SE = 0.55; 95% CI [0.41, 2.57]) indicates that the ratio of proportions had a greater effect for *target equal* trials compared to *total equal* trials. Exponentiating the model coefficients for main effects revealed that incremental increases in the ratio of proportions (β_{RP} = 0.09; SE = 0.05; 95% CI [-0.01, 0.18) led to small increases in the odds of making a correct choice and *total equal* trials increased the odds of a correct choice compared to *target equal* trials (β_{TT} = -1.38; SE = 0.87; 95% CI [-3.08, 0.32). Similar comparisons were conducted for the LMMs used to analyze reaction time data, the details of which can be found in the Supplemental Material.

Correlation with number approximation acuity. Due to experimenter error, number approximation data from the Panamath task was only collected from 30 of the 48 participants. Correlational analyses revealed that Weber Fraction as measured by the Panamath task was significantly negatively correlated with overall performance in the probability task for this subsample of participants (Pearson's $r = -.38$, 95% CI $[-.65, -.02]$, $t(28) = -2.15$, $p = .040$). This indicates that participants with better numerical acuity (lower Weber fractions) performed better on the probability task.

2.2.3 Discussion

Results from Experiment 1 revealed that as the ratio of proportions increased, the two proportions become easier to discriminate. Furthermore, participant's number sense acuity was correlated with performance on the probability judgment task (i.e. lower Weber fractions as measured by Panamath predicted more accurate probability estimation). Finally, participants performed worse on *target equal* trials compared to *total equal* trials, particularly when the ratio of proportions was small, suggesting that participants may have relied on a decision-making heuristic such as 'pick the group with more target marbles' or 'pick the group with fewer non-target marbles'.

However, two important issues in the design of the experiment need to be resolved before any strong conclusions can be made. Although performance was significantly lower for *target equal* trials compared to *total equal* trials, it is impossible to identify whether this was due to

participants focusing on the number of target objects or the number of non-target objects. Additionally, if probability estimation relies on number approximation, then the same non-numerical features which influence performance on number approximation tasks should also influence performance on the probability judgment task. Experiment 2 was undertaken in order to replicate the main results from Experiment 1 and to extend these findings to account for non-numerical features that were not previously controlled for. If number approximation abilities support probability estimation, then perceived probability should be influenced by the same numerical and non-numerical features known to influence perceived numerosity.

2.3 Experiment 2

In Experiment 2, adult participants recruited online were asked to perform similar probability approximation and dot approximation tasks as those reported in Experiment 1. We addressed the limitation of Experiment 1 by incorporating three trial types meant to investigate the influence of the number-based heuristic, 'pick the group with the most [target color] marbles' as well as the area-based heuristic, 'pick the group with the greatest [target color] area'. In addition we modified a model of number approximation from the psychophysical literature on perceived numerosity to account for numerical and non-numerical stimulus features in probability judgments.

2.3.1 Psychophysical Model of number approximation extended to ratio

Recently, DeWind et al. (2015) proposed a model to account for both numerical and non-numerical stimulus features in data from a dot discrimination task. By systematically varying the size, spacing, and number of dots, and recording participants' numerosity judgments, DeWind and colleagues were able calculate a participant's 'bias' for both the size of individual and collective objects as well as the sparsity of the visual array as measured by the ratio of the number of objects to the field area containing all of the objects of the visual array. In the current paper, we adapt this model to account for the ratios of the proportions of marbles presented in each image in order to assess the psychophysical properties of probability estimation.

Following the methods outlined in DeWind et al. (2015), coefficients for size (β_{Size}), spacing (β_{Space}) and number (in this case proportion, $\beta_{Proportion}$) were generated using non-linear regression, where the participant's response is regressed over the log of the ratio of proportions of marbles, the ratio of the proportions of the sizes of the marbles and the ratio of proportions of the sparsity of each array of marbles. Model predictions were generated by entering the coefficients into the following equation,

$$p(CorrectChoice) = (1-\gamma)\left(\frac{1}{2}\left[1+erf\left(\frac{log2(RP_{Proportion}) - \left(\frac{-\beta_{Side} - \beta_{Size}log_2(RP_{Size}) - \beta_{Space}log_2(RP_{Space})}{\beta_{Proportion}}\right)}{\sqrt{2}\frac{1}{\beta_{Proportion}}}\right)\right] - \frac{1}{2}\right) + \frac{1}{2} \quad [1]$$

where γ is a guessing parameter, RP_{num}, RP_{size}, RP_{space} refer to the ratios of the proportions of marbles, size and spacing of marbles being approximated, β coefficients represent the 'decision weight' induced by the ratio of the proportions ($\beta proportion$), placement of marbles

on the screen ($\beta side$), the ratio of the size of the surface area ($\beta size$) and the ratio of the field areas of the two arrays ($\beta space$) and erf is the mathematical error function.

Importantly, this equation is the same as that presented in DeWind et al. (2015) with one notable exception. As a model of number approximation, numerical and non-numerical stimulus features in the original model were represented as the ratio of the stimulus features of the two arrays of dots. In order to model probability estimation, the variables for numerical and non-numerical stimulus features in the current model ($RP_{Proportion}$, RP_{Size}, and RP_{Space}) are represented as ratios of proportions of the stimulus features.

2.3.2 Methods

Participants. Eighty adult participants were recruited online from Amazon's Mechanical Turk via PsiTurk (Gureckis et al., 2016), an open source platform for online behavioral experiments. Participants completed an online consent form approved by the [university IRB information redacted for blind review]. Since we decided to collect data from an online sample we were concerned about dropout rate and inattention thus we doubled our target sample from Experiment 1. Seven participants were excluded for not passing practice trials. Practice trials in the probability task consisted of one bin containing all red marbles and another bin containing all white marbles. In order to be included in the final sample, participants had to choose the bin containing the color of marbles that they were instructed to collect on 3 out of the 4 practice trials. Ten participants failed to reach inclusion criteria during either the probability task practice trials, the dot approximation practice trials or both leaving a final sample of N = 70 (Mean age = 34.18, SD = 8.83; 42 females).

Materials. As with Experiment 1, images were rendered using Blender 2.72, with some key differences. First, the location of each marble was randomly generated for each image. Second, three trial types were created for each of the ratios of proportions presented in Table 2 below. *Total equal* trials consisted of groups with the same total number or marbles ranging from 10 to 40 marbles. *Number vs proportion* trials presented the same proportions as *total equal* trials except that the number of target marbles in the 'losing' distribution was larger than the number of target marbles in the 'winning' distribution. Finally, area coefficients were applied to the size of the marbles in *area-anticorrelated* trials such that the percentage of the area of the target marbles (e.g., red) in the 'losing' distribution matched the numerical proportion of the target marbles to total marbles in the 'winning' distribution. A similar practice was used in Halberda and Feigenson (2008) in order to control for the effects of area in number approximation tasks. Eight images were created for each ratio of proportions and trial type resulting in a total of 528 images. Images were divided equally into four conditions based on target color and the order of presentation of the images. Each participant was randomly assigned to one of four conditions (Red, order 1; Red, order 2; White, order 1; and White, order 2) and viewed 132 images.

Table 2.2 Proportions presented in each trial of Experiment 2.

Proportion group 1	Proportion group 2	Ratio of Proportions
0.55	0.50	1.10
0.50	0.45	1.11
0.45	0.40	1.12
0.55	0.45	1.22
0.80	0.60	1.33
0.80	0.55	1.45
0.60	0.40	1.50
0.75	0.50	1.50
0.95	0.55	1.73
0.95	0.50	1.90
0.50	0.25	2.00
0.40	0.15	2.67
0.60	0.20	3.00
0.50	0.15	3.33
0.80	0.20	4.00
0.40	0.10	4.00
0.75	0.15	5.00
0.95	0.15	6.33
0.80	0.10	8.00
0.85	0.10	8.50
0.90	0.10	9.00
0.95	0.10	9.50

Note. Ratios of Proportions are rounded to 2 digits. See the Supplemental Material for a full table including the numbers of marbles used in each group for each trial type.

Procedure. After providing informed consent, participants were forwarded to a secure PsiTurk server for the experiment. Participants first viewed an instruction screen describing the probability judgment task. After reading through the instructions, participants were presented with practice trials in order to ensure that they understood the instructions. Following the practice trials, participants were presented with 132 semi-randomized test trials in which they were able to view the images for 750 ms. As with Experiment 1, test trials were semi-randomized in order to reduce the possibility of a participant inadvertently learning an incorrect choice rule. After each set of 25 test trials, participants were encouraged to take a short break if needed. Figure 3 below provides a visual schematic of the procedure including an example of each trial type.

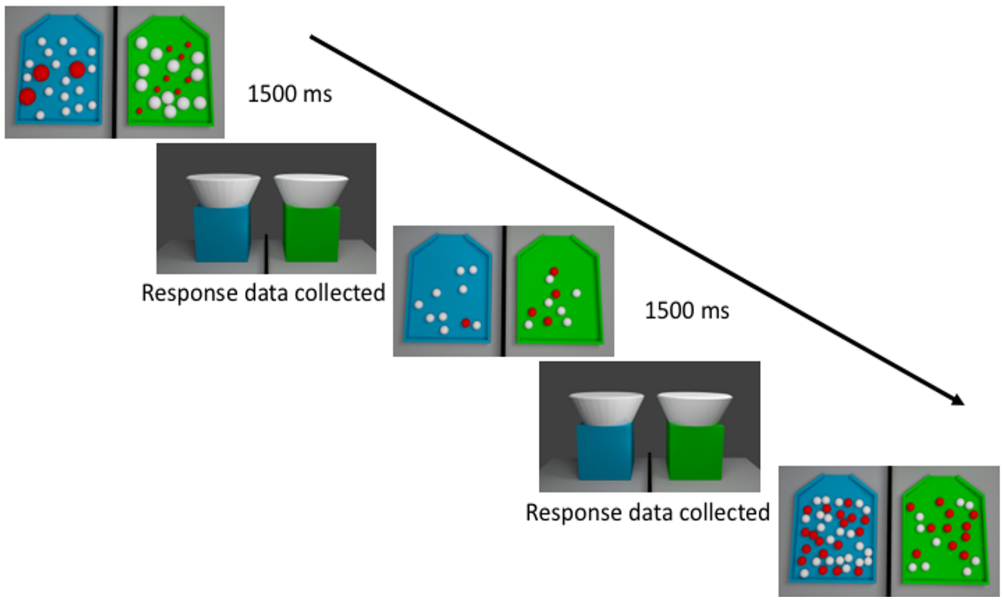

Figure 2.3 Diagram of the experimental procedure used in Experiment 2. The sample image at the top presents an area-anticorrelated trial, the sample image in the middle presents a total equal trial and the sample image at the bottom presents a number vs proportion trial.

Once the participant completed the probability judgment test trials they were informed that they would perform a second task which was similar to the first but different in some very important ways. Due to the structure of online experiments Panamath was not used as the dot approximation task in Experiment 2. A similar task allowing for more experimental control over stimuli and timing was used (Odic, 2018). After reading the instructions for the dot approximation task, participants were presented with two practice trials meant to ensure that they understood the instructions.

2.3.3 Results

General accuracy. Participants performed significantly above chance in all three trial types (*total equal*: $M = 0.88$, $SD = 0.11$; $M = 0.88$, 95% CI [0.85, 0.91], $t(69) = 29.26$, $p < .001$; *number vs proportion*: $M = 0.69$, $SD = 0.19$; $M = 0.69$, 95% CI [0.64, 0.74], $t(69) = 8.17$, $p < .001$; *area-anticorrelated*: $M = 0.70$, $SD = 0.18$; $M = 0.70$, 95% CI [0.66, 0.75], $t(69) = 9.36$, $p < .001$).

Statistical models. Results of GLMM comparisons revealed that the model with the best fit to the data predicted the participant's response from ratio of proportions, trial type and the interaction between ratio of proportions and trial type (AIC_{RP*TT} =7,774.30). This model outperformed the null model (AIC_{null} = 9,087.24; χ^2= 1,322.94; df = 5; $p < .001$) and the models predicting performance from the ratio of proportions alone (AIC_{RP} = 8,247.66; χ^2= 481.36; df = 4; $p < .001$), trial type alone (AIC_{TT} = 8,675.46; χ^2= 907.16; df = 3; $p < .001$), as well as the model predicting performance from both ratio of proportions and trial type with no interactions (AIC_{RP+TT} = 7,786.80; χ^2= 16.50; df = 2; $p = .028$).

Inspection of model coefficients revealed that an increase in the ratio of proportions (β_{RP} = 0.44; SE = 0.04; 95% CI [0.35, 0.52]) led to an increase in the odds of a correct choice on *total equal* trials ($\beta_{Intercept}$ = 1.04; SE = 0.14; 95% CI [0.76, 1.33]). The odds of responding with a correct choice decreased for both *number vs proportion* trials (β_{NvP} = -1.43; SE = 0.13; 95% CI [-1.70, -1.17]) and *area-anticorrelated* trials (β_{AA} = -1.04; SE = 0.13; 95% CI [-1.30, -0.78]) relative to *total equal* trials. The model coefficients for the interaction of ratio of proportions and trial type indicated that the effect of ratio of proportions was reduced for *area-anticorrelated* trials (β_{RP*AA} = -0.13; SE = 0.05; 95% CI [-0.22, -0.03]) relative to *total equal* trials, but there was no effect of ratio of proportions on *number vs proportion* trials (β_{RP*NvP} = -0.01; SE = 0.05; 95% CI [-0.11, 0.09]). Linear regression analyses of reaction time data yielded similar results (see supplemental material).

Correlation with number approximation acuity. As in Experiment 1, we find that number sense acuity as measured by the dot approximation task was significantly correlated with overall performance in the probability task (Pearson's $r = -.43$, 95% CI $[-.60, -.22]$, $t(68) = -3.94$, $p < .001$). Figure 4 displays average performance by ratio of proportions and trial type.

Psychophysical model of number approximation extended to ratio. Model predictions were generated for each trial image presented to participants by entering each participant's model coefficients as well as the values for the stimulus features for each trial image into Equation 1. Figure 4 presents the model predictions alongside the data from Experiment 2. Importantly, model predictions were significantly positively correlated with response data ($r = .42$, 95% CI $[.41, .44]$, $t(9{,}238) = 44.94$, $p < .001$). Results of t-tests revealed that all β coefficients were significantly greater than 0 ($\beta proportion$: $M = 1.75$, 95% CI [1.26, 2.24], $t(69) = 7.12$, $p < .001$; $\beta size$:$M = 0.31$, 95% CI [0.04, 0.57], $t(69) = 2.32$, $p = .023$; $\beta space$: $M = 0.14$, 95% CI [0.02, 0.26], $t(69) = 2.25$, $p = .028$), except for the $\beta side$ coefficient ($M = -0.04$, 95% CI $[-0.18, 0.11]$, $t(69) = -0.52$, $p = .606$). These results provide further evidence that probability estimation share similar psychophysical properties as those reported for the approximate number sense. Furthermore, the $\beta proportion$ coefficient was significantly greater than all other model coefficients ($\beta space$: $M_d = 1.62$, 95% CI [1.07, 2.16], $t(69) = 5.93$, $p < .001$; $\beta size$: $M_d = 1.44$, 95% CI [0.81, 2.08], $t(69) = 4.54$, $p < .001$; $\beta side$: $M_d = 1.79$, 95% CI [1.26, 2.32], $t(69) = 6.76$, $p < .001$) indicating that although participants' choices were influenced by non-numerical stimulus features, they relied most on the numerical proportions of the marbles in each group when making their decisions.

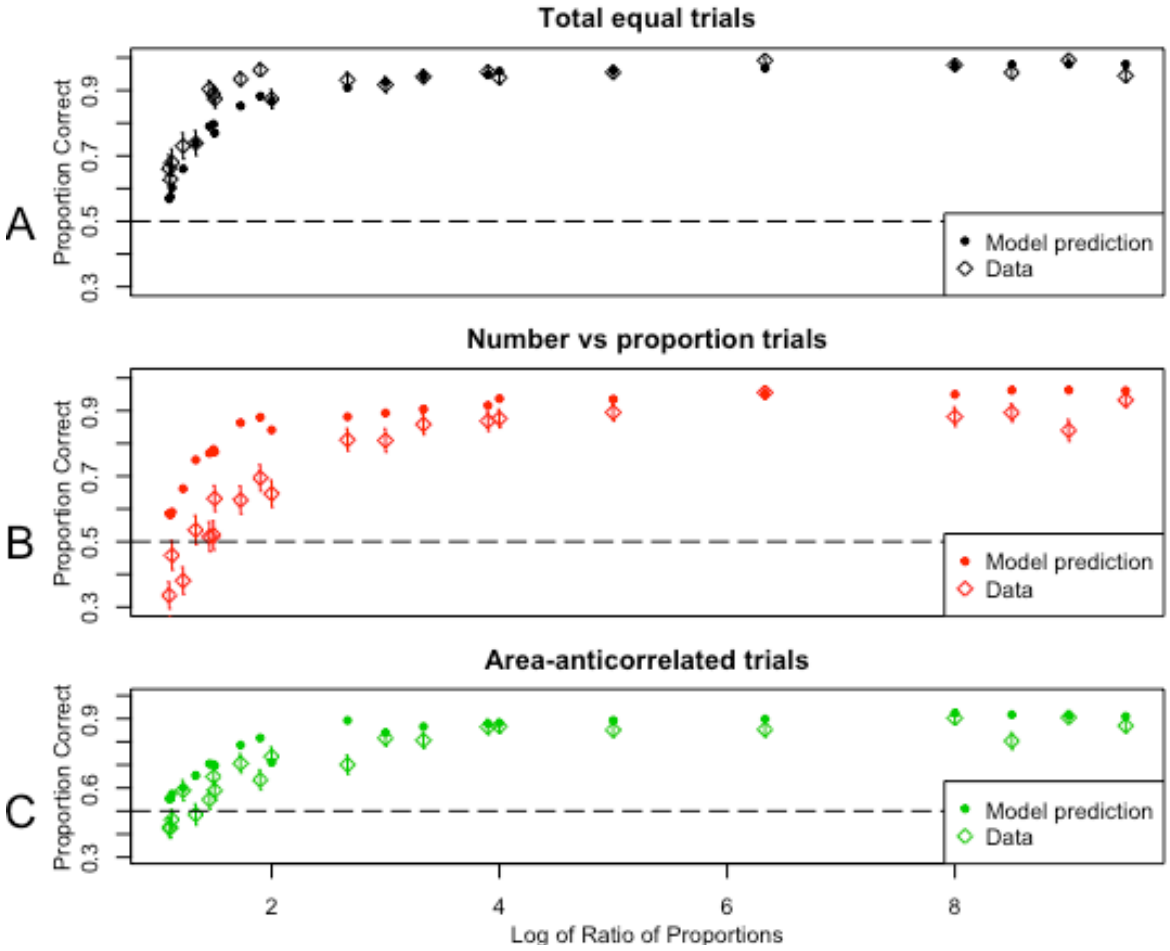

Figure 2.4 Model predictions alongside the average proportion of correct responses by ratio of proportions in Experiment 2. A) *Total equal* trials. B) *Number vs proportion* trials. C) *Area-anticorrelated* trials. Error bars indicate bootstrapped 95% Confidence Intervals

2.3.3 Discussion

Experiment 2 offers a direct replication and extension of Experiment 1. As in Experiment 1 we found that the number of target marbles influences participants' decisions. Furthermore, performance on *number vs proportion* trials compared to *target equal* trials suggests that participants demonstrated a bias toward groups with greater amounts of target marbles. Results from Experiment 2 also revealed that the area of target marbles influenced performance suggesting that probability estimation is affected by the same numerical and non-numerical stimulus features that influence perceived numerosity. Inspection of model coefficients revealed an interaction between ratio of proportions and *area-anticorrelated* trial types indicating that participants made fewer errors on area anticorrelated trials with greater ratios of proportions (i.e. greater distance between the two compared proportions). When the probabilities were more difficult to discriminate, participants were biased to choose the array with larger target marbles. This suggests that although participants were primarily choosing based on proportion, they were more biased by the (irrelevant) size of the marbles when the ratio was very difficult to discriminate. Furthermore, results of analyses of the coefficients for the psychophysical model

suggest that the ratio of proportions had the strongest effect on the participants' decisions, but that the size and spacing of the marbles also contributed to their decision.

Together the results of Experiments 1 and 2 suggest that both number approximation and heuristic decision rules play a role in probability estimation based on proportions. In the context of previous research on probability approximation abilities of both children and adults (Eckert et al., 2018; Fazio et al., 2014; O'Grady & Xu, 2019, O'Grady et al. 2016), as well as the literature on probabilistic reasoning based on exact, countable quantities (Alonso & Fernández-Berrocal, 2003; Denes-Raj & Epstein, 1994; Falk et al., 2012; Kirkpatrick & Epstein, 1992; Pacini & Epstein, 1999; Piaget & Inhelder, 1975), the current findings seem to indicate that heuristic decision rules such as 'pick the group with the largest number of target outcomes' are quite common in probability estimation based on proportion, particularly when task is difficult, in both adults and children. However, in Experiments 1 and 2 it is possible that the bias toward choices with a greater number of target marbles is the result of the brief presentation time (750ms). Although there is theoretically enough time to approximate the different sets of marbles presented in the images (Halberda et al., 2006), it is possible that the short presentation time caused participants to switch to a faster yet less accurate decision-making strategy. Indeed, researchers studying number approximation abilities have found that people make more accurate approximations of number when they are provided with greater presentation time (Inglis & Gilmore, 2013). In Experiment 3 we address this question by presenting the same images used in Experiment 2 during counterbalanced blocks with short (750ms) and long (1500ms) presentation times.

2.4 Experiment 3

In Experiment 3, we attempt to investigate the stability of the use of heuristic rules with respect to stimulus presentation time. Participants were recruited online to perform the same task as that reported in Experiment 2 except that they played the game twice. If the use of the heuristic decision rules 'pick the group with the most [target color] marbles' or 'pick the group with the greatest [target color] area' is related to the short stimulus presentation time used in Experiment 2, then we should find an interaction between trial type and presentation time in the current experiment such that participants rely less on such heuristics when the presentation time is longer.

2.4.1 Methods

Participants. Results from Experiment 2 revealed that only 7 of 80 participants failed the practice trials and for this reason we returned to the sample size used in Experiment 1. Forty participants were recruited for an online decision-making study via Amazon's Mechanical Turk. We excluded the data from 7 participants because they failed to choose correctly on at least 75% (6/8) of the practice trials. Our final sample consisted of N = 33 English speaking participants (Mean Age = 35.62; SD = 10.65; 13 females). Due to a programming error, age data were not collected from half of the participants.

Material. We used the same set of images and procedure described in Experiment 2 with one critical difference. Each participant performed the entire task twice in two different blocks of 132 trials. In one block, participants viewed the images for 750 ms following the procedure outlined in Experiment 2. In the other block, the same images were presented for 1500 ms, thus doubling the time that participants had to approximate the probabilities and make their decision. The order in which participants received the blocks was counterbalanced.

2.4.2 Results

Generalized Linear Mixed Effects Regression Analyses of performance. We used the same analytical methods reported in Experiment 2. Comparisons revealed that the model with the best fit to the data predicted the binary response variable based on ratio of proportions and trial type with no interaction between the two variables (AIC_{RP+TT} =6,248.82). This model outperformed the null model (AIC_{Null} =6,918.47; χ^2= 675.65; *df* = 3; $p < .001$), and the models predicting performance from the ratio of proportions alone (AIC_{RP} = 6,459.80; χ^2= 214.98; *df* = 2; $p < .001$), trial type alone (AIC_{TT} = 6,721.24; χ^2= 474.42; *df* = 1; $p < .001$), as well as the model for the interaction between the two variables (AIC_{RP+TT} = 6,250.70; χ^2= 2.12; *df* = 2; *p* 0.35). Analyses of reaction time data yielded similar results (see Supplemental Material).

Inspection of the model coefficients reveals that participants once again performed above chance ($\beta_{Intercept}$ = 1.62; SE = 0.19; 95% CI [1.25, 1.99, $p < .001$]) and performance improved as the ratio of the proportions increased (β_{RP} = 0.30; SE = 0.02; 95% CI [0.27, 0.33, $p < .001$]). As with Experiment 2, analyses revealed main effects for both *number vs proportion* (β_{NvP} = -1.08; SE = 0.09; 95% CI [-1.24, -0.91, $p < .001$]) and *area anticorrelated* trial types (β_{AA} = -1.06; SE = 0.09; 95% CI [-1.23, -0.89, $p < .001$]). However, the interaction between ratio of proportions and trial type reported in Experiments 1 & 2 was not replicated in the current experiment.

Analyses of presentation time. Analyses of presentation time and block order revealed that the model with the best fit to the data predicted the binary response variable based on presentation time and block order without an interaction between presentation time and block order (AIC_{PT+BO} =6,823.35). This model outperformed then null model (AIC_{Null} =6,918.47; χ^2= 99.11; *df* = 2; $p < .001$) as well as the models predicting performance based on presentation time (AIC_{PT} =6,874.10; χ^2= 52.75; *df* = 1; $p < .001$) and block order (AIC_{BO} =6,847.08; χ^2= 25.73; *df* = 1; $p = .00$) alone. Importantly, the model predicting performance based on the interaction between presentation time and block order was not a significantly better fit to the data (AIC_{PT*BO} =6,824.82; χ^2= 0.53; *df* = 1; $p = .47$) thus the simpler model without an interaction is preferred to the more complex model since it can explain the same amount of variance with fewer parameters.

Compared to the intercept of the model ($\beta_{intercept}$ = 1.37; SE = 0.17; 95% CI [1.04, 1.69) which represents the odds of a correct response when the short presentation time (750ms) occurred in the first block of trials, the longer presentation time (β_{1500ms} = 0.32; SE = 0.06; 95% CI [0.20, 0.44) led to greater improvements in performance. Additionally, we find that performance was also higher for the trials in the second block compared to trials in the first block (β_{Block2} = 0.46; SE = 0.06; 95% CI [0.34, 0.58). This finding suggests that general familiarity with the task improves performance, even in the absence of explicit feedback.

Interaction with trial type. In order to investigate strategy use in the current task, we assessed whether presentation time influenced the effect of the trial type. Our results revealed that the model with the best fit to the data predicted the correct binary response variable based on presentation time and trial type (AIC_{PT+TT} =6,675.61). This model outpredicted the null model (AIC_{Null} =6,918.47; χ^2= 248.85; *df* = 3; $p < .001$), as well as the models for presentation time alone (AIC_{PT} =6,874.10; χ^2= 202.49; *df* = 2; $p < .001$) and trial type alone (AIC_{TT} =6,721.24; χ^2= 47.63; *df* = 1; $p < .001$). The model for main effects did not outperform the model with an interaction between trial type and presentation time (AIC_{PT*TT} =6,675.33; χ^2= 4.28; *df* = 2; p .12). Since the $PT + TT$ model has fewer parameters than the $PT * TT$ model, it is considered the superior model as it predicted the same amount of variance with fewer degrees of freedom. These results indicate that presentation time did not influence the strategy use because performance improved for all three trial types rather than only improving for the *number vs proportion* trials.

Inspection of the model coefficients ($\beta_{intercept}$ = 2.27; SE = 0.18; 95% CI [1.93, 2.62, $p <$ .001) revealed that the longer presentation time (β_{1500ms} = 0.43; SE = 0.06; 95% CI [0.31, 0.55, $p < .001$) had a positive effect on performance while both *number vs proportion* trials (β_{NvP} = -1.01; SE = 0.08; 95% CI [-1.17, -0.85, $p < .001$) and *area anticorrelated* trials (β_{AA} = -1.00; SE = 0.08; 95% CI [-1.16, -0.84, $p < .001$) had a negative effect on performance relative to *total equal* trials. Figure 5 presents the average accuracy by ratio of proportions and presentation time for *total equal* trials (A), *number vs proportion* trials (B), and *area anticorrelated* trials (C).

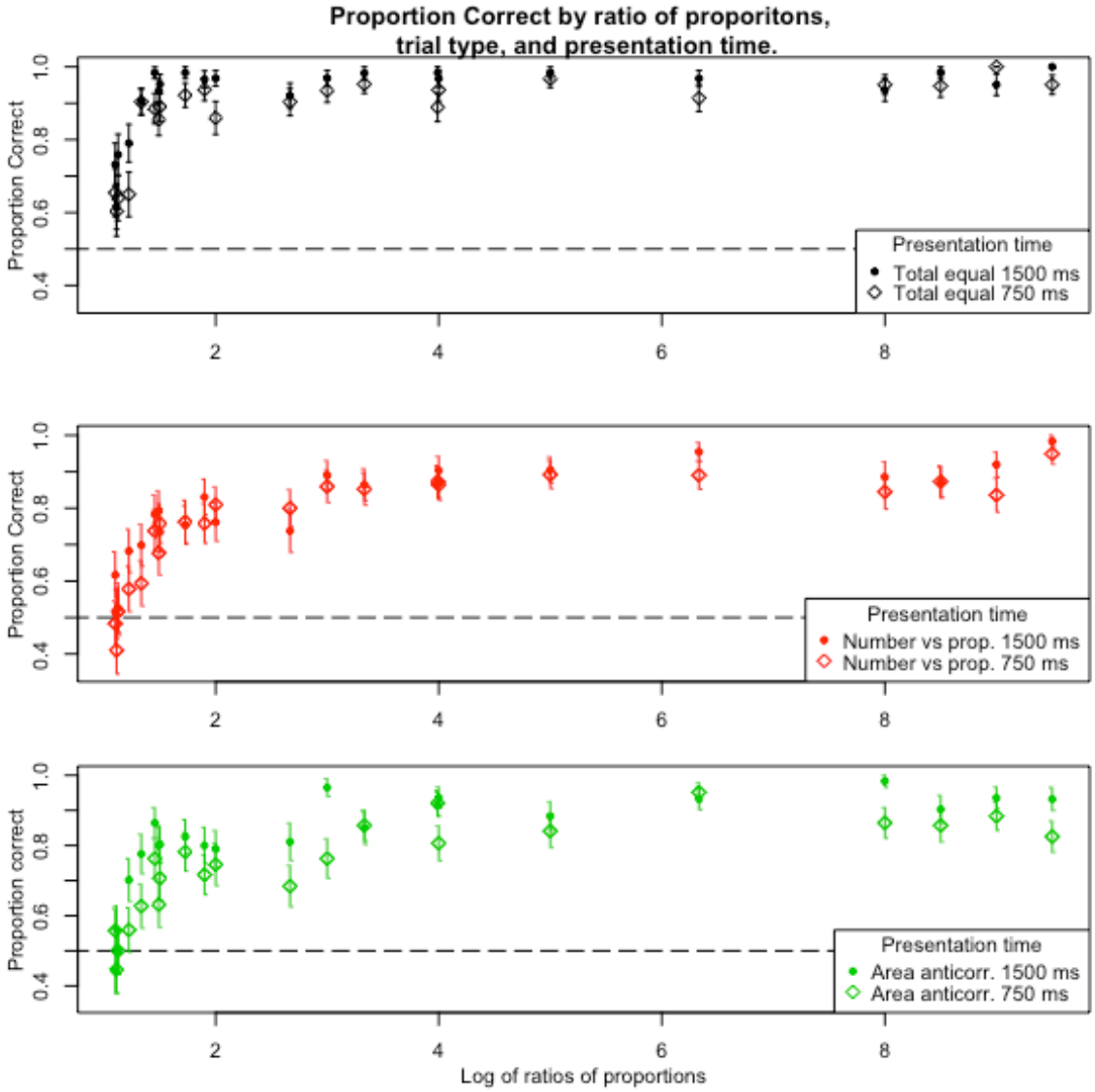

Figure 2.5 Proportion correct responses by ratio of proportions, presentation time, and trial type. Trial type is represented as color with black indicating *total equal* trials, red indicating *number vs* proportion trials and green indicating *area-anticorrelated* trials. Solid dots indicate data from the long stimulus presentation time while diamonds indicate data from the short presentation time. Error bars indicate bootstrapped 95% Confidence Intervals.

2.4.3 Discussion

In Experiment 3 we replicated the main effect reported in Experiment 2, providing even more evidence that the ability to discriminate probability is influenced by the perceived magnitude of the proportions being estimated as well as the absolute number of the target outcomes and relative size of the enumerated objects. Furthermore, results revealed that participants once again performed worse on *number vs proportion* trials compared to *total equal* trials suggesting the use of heuristic decision biases such as 'pick the group with the greater number of target color

marbles'. By manipulating the stimulus presentation time, we also investigated whether presentation time influences the use of heuristic strategies. Results of the GLMM comparisons revealed that although the longer presentation time resulted in improved accuracy, this effect did not interact with trial type, which suggests that participants used the same decision strategy across both presentation times.

There are two possible explanations for this heuristic in probability estimation. The first is that adults employ simple heuristics in a resource-rational way (Griffiths, Lieder, & Goodman, 2015). That is, when people's ability to compute probability is constrained by time (i.e. 750 – 1500 ms), they fall back on simple magnitude information such as area and absolute number. However, an alternative explanation is that participants in our specific task believed that they would not encounter choices involving groups with different total amounts of marbles. In Experiment 4, we provided explicit instructions meant to reduce the use of this incorrect strategy.

2.5 Experiment 4

In Experiment 4 we presented online participants with the same task used in Experiment 3 except that before the first block of trials participants read an additional instruction screen informing them that there is always a correct choice and that on some trials the correct choice has more [target color] marbles but on other trials the correct choice has fewer [target color] marbles. If participants simply did not understand the purpose of the task in Experiment 3 then their performance should improve after viewing the instruction screen.

2.5.1 Methods

Participants. As with Experiment 3, we recruited 40 adult participants online via Amazon's Mechanical Turk. Data from 3 participants were excluded because they failed to choose correctly on at least 75% (6/8) of the practice trials. Our final sample consisted of N = 37 English speaking participants (Mean age = 36.01; SD = 12.02; 12 females). Due to a programming error, age data were not collected for the 10 participants reported in Experiment 4, and one participant declined to provide age data.

Procedure. We used the same methods as Experiment 3 except that participants viewed an additional instruction screen before they performed the 1st block of trials. This additional instruction screen presented the following instructional hint: "There is always a correct choice in this game. Sometimes the correct choice has more [target color] marbles but at other times the correct choice has fewer [target color] marbles compared to the incorrect choice." All 40 participants received these additional instructions before their first block thus half of the participants received these instructions before the block of trials with a short presentation time (750 ms) while the other half received the instructions before the block of trials with a long presentation time (1500 ms).

2.5.2 Results

We first present the results of analyses of data from Experiment 4 in order to ensure that our main findings from Experiments 2 & 3 were replicated in the current sample. Next, we combine these data with those from Experiment 3 in order to test the effect of instruction on participants' choice strategy.

Generalized Linear Mixed Effects Regression Analyses of performance. We used the same analytical methods reported in Experiments 2 and 3. Model comparisons for the predicted variables of interest revealed that the model with the best fit to the data predicted the binary response variable based on ratio or proportions, trial type, and the interaction between the two variables (AIC_{RP*TT} =7,576.28). This model outperformed the null model (AIC_{Null} =8,421.34; χ^2= 855.06; *df* = 5; $p < .001$), and the models predicting performance from the ratio of proportions alone (AIC_{RP} = 7,882.21; χ^2= 313.93; *df* = 4; $p < .001$), trial type alone (AIC_{TT} = 8,148.73; χ^2= 578.45; *df* = 3; $p < .001$) as well as the combined model with trial type and ratio of proportions without any interactions (AIC_{RP+TT} = 7,589.33; χ^2= 17.05; *df* = 2; $p < .001$). Analyses of reaction time data yielded similar results (see Supplemental Material).

Inspection of the model coefficients reveals that participants once again performed above chance ($\beta_{Intercept}$ = 1.32; SE = 0.19; 95% CI [0.96, 1.68, $p < .001$]) and performance improved as the ratio of the proportions increased (β_{RP} = 0.43; SE = 0.04; 95% CI [0.34, 0.51, $p < .001$]). As with Experiment 2, analyses revealed main effects for both *number vs proportion* (β_{NvP} = -0.95; SE = 0.13; 95% CI [-1.22, -0.69, $p < .001$]) and *area anticorrelated* trial types (β_{AA} = -0.59; SE = 0.13; 95% CI [-0.86, -0.33, $p < .001$]). We also found an interaction between ratio of proportions and trial types indicating that the effect of ratio of proportions was slightly decreased for both *number vs proportion* (β_{RP*NvP} = -0.12; SE = 0.05; 95% CI [-0.22, -0.03, p .01]) and *area anticorrelated* (β_{RP*AA} = -0.18; SE = 0.05; 95% CI [-0.28, -0.09, $p < .001$) trial types relative to *total equal* trials.

Analyses of presentation time and block order. Analyses revealed the same pattern reported in Experiment 3. The model with the best fit to the data predicted the binary response variable from presentation time and block order without an interaction between the two variables (AIC_{PT+BO} =8,374.78). This model outperformed the null model (AIC_{Null} =8,421.34; χ^2= 50.56; *df* = 2; $p < .001$) as well as the models predicting performance based on presentation time (AIC_{PT} =8,385.58; χ^2= 12.80; *df* = 1; $p < .001$) and block order (AIC_{BO} =8,409.08; χ^2= 36.29; *df* = 1; p = .00) alone. Importantly, the model predicting performance based on the interaction between presentation time and block order was not a significantly better fit to the data (AIC_{PT*BO} =8,375.99; χ^2= 0.79; *df* = 1; p = .37) as with previous reports the simpler model is preferred to the more complex model as it can explain the same amount of variance with fewer parameters.

Compared to the short (750 ms) presentation time ($\beta_{intercept}$ = 1.41; SE = 0.14; 95% CI [1.13, 1.69), the longer presentation time (β_{1500ms} = 0.34; SE = 0.06; 95% CI [0.23, 0.45) led to greater performance. Additionally, we find performance improved in the second block compared to the first block (β_{Block2} = 0.20; SE = 0.06; 95% CI [0.09, 0.31).

Interaction with trial type. In order to investigate strategy use in the current task, we assessed whether presentation time influenced the effect of the trial variables reported in previous studies and analyses. Our results revealed that the model with the best fit to the data predicted the correct binary response variable based on presentation time and trial type (AIC_{PT+TT} =8,111.32). This model outpredicted the null model (AIC_{Null} =8,421.34; χ^2= 316.03; *df* = 3; $p < .001$), as well as the models for presentation time alone (AIC_{PT} =8,385.58; χ^2= 278.27; *df* = 2; $p < .001$) and trial type alone (AIC_{TT} =8,148.73; χ^2= 39.42; *df* = 1; $p < .001$). The model for main effects did not outperform the model with an interaction between trial type and presentation time (AIC_{PT*TT} =8,114.14; χ^2= 1.18; *df* = 2; $p = .56$). Since the $PT + TT$ model has fewer parameters than the $PT * TT$ model, it is considered the superior model as it predicted the same amount of variance with fewer degrees of freedom. These results again indicate that presentation time did not influence the strategy use because performance improved equally for all three trial types in the longer presentation time condition compared to the shorter presentation time condition.

Inspection of the model coefficients ($\beta_{intercept}$ = 2.29; SE = 0.15; 95% CI [1.98, 2.59, $p < .001$) revealed that the longer presentation time (β_{1500ms} = 0.36; SE = 0.06; 95% CI [0.24, 0.47, $p < .001$) had a positive effect on performance while both *number vs proportion* trials (β_{NvP} = -1.15; SE = 0.08; 95% CI [-1.30, -1.00, $p < .001$) and *area anticorrelated* trials (β_{AA} = -0.97; SE = 0.08; 95% CI [-1.12, -0.82, $p < .001$) had a negative effect on performance relative to *total equal trials*. Figure 6 presents the average accuracy by ratio of proportions and presentation time for *total equal* trials (A), *number vs proportion* trials (B), and *area anticorrelated* trials (C).

Interaction between trial type and instructions. We combined the data from Experiments 3 & 4 in order to investigate the effect of instructions on performance. Results of the comparisons of models predicting the binary response variable from trial type, instructions, and the interaction between instructions and trial type revealed that the model with the best fit predicted the response variable from trial type alone (AIC_{TT} =14,865.62). This model outperformed the null model (AIC_{Null} =15,336.21; χ^2= 316.03; *df* = 3; $p < .001$) as well as the model for instructions alone ($AIC_{Instructions}$ =15,338.11). Furthermore, the more complex models involving trial type and instructions ($AIC_{Instructions+TT}$ =14,867.53; χ^2= 0.10; *df* = 1; $p = .76$) and the interaction between the two variables ($AIC_{Instructions*TT}$ =14,868.22; χ^2= 3.41; *df* = 3; $p = .33$) did not outperform the simpler model. Therefore, the data suggest that task instructions did not significantly influence participants' performance. Figure 6 plots the average performance for experiments 3 and 4 by trial type.

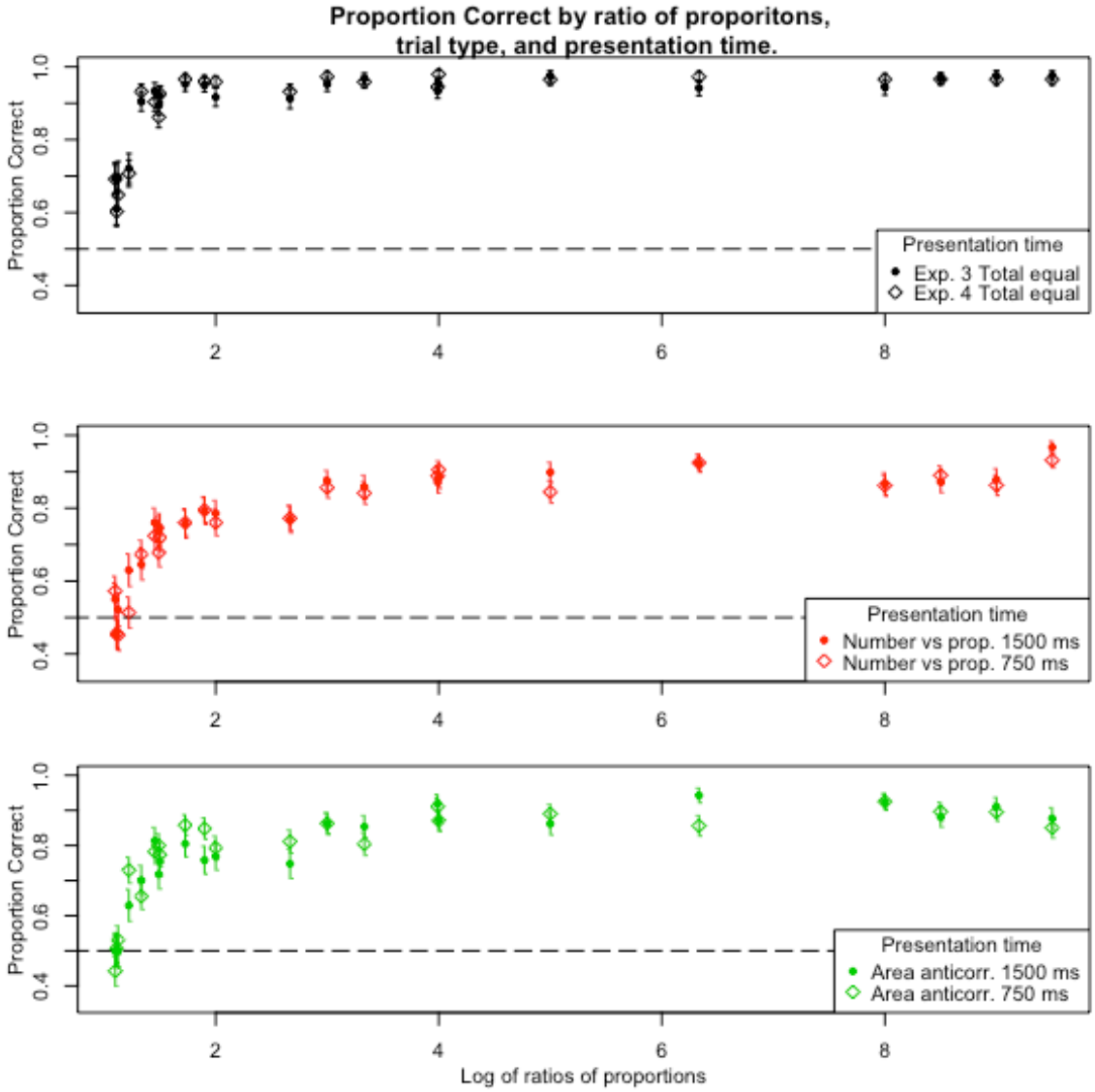

Figure 2.6 Proportion correct responses by ratio of proportions, trial type, and experiment. Trial type is represented as color with black indicating *total equal* trials, red indicating *number vs* proportion trials and green indicating *area-anticorrelated* trials. Solid dots indicate data from Experiment 3 (no instructions) while diamonds indicate data from Experiment 4 (instructions). Error bars indicate bootstrapped 95% Confidence Intervals.

2.5.3 Discussion

Results from Experiment 4 suggest that instructions alone could not reduce the bias toward choices with a larger number of target color marbles in the current task. When the combined data from Experiments 3 and 4 were analyzed to investigate the effect of instructions and trial type, model comparisons revealed that the superior model predict participant responses from trial type alone. Therefore, the heuristic of choosing the array with the larger number of target marbles is not merely a byproduct of participants' expectations about the task.

2.6 General Discussion

In four experiments we report several important findings on the mental representation of non-symbolic probability. A critical innovation of the current research is that we incorporated different trial types in order to tease apart the influence of non-numerical stimulus features and heuristic biases in human probability estimation. In Experiment 1 we show that adults can accurately make rapid probability estimations based on the proportion of discrete events. Furthermore, performance on this task is correlated with the acuity of the participants' number sense, indicating that number approximation and probabilistic discrimination rely on related numerical processing abilities. In Experiment 2 we replicate and extend these findings to show that these estimations are also influenced by the same numerical and non-numerical stimulus features that influence perceived numerosity, providing further evidence that the two approximation tasks rely on the same analog magnitude processing system. Importantly, in both experiments we have shown that human probability estimation is influenced by the perceived numerosity of target events (a formally incorrect strategy for estimating probability) rather than the actual proportion of target events. In Experiment 3, we show that this decision bias persists even when participants are given more time to make their decisions. Furthermore, in Experiment 4 we show that this bias is also not the result of a failure to understand the task. Together, the results of Experiments 3 & 4 suggest that the bias towards selecting groups with a larger number of target elements in probability estimations may not represent an incorrect understanding of the proportional probability but rather represents a crucial component of the human decision-making process: when probabilities are easily discriminated, people can accurately rely on ratio processing; but, as indicated by the interaction between ratio of proportions and trial type in Experiments 1, 2 & 4, when the magnitude of two probabilities are close together, adults rely on heuristic decision rules.

While previous research has provided evidence of the role of analog magnitude processing in probability estimation (Eckert et al., 2018; O'Grady & Xu, 2019), and revealed a correlation between number approximation acuity and proportional reasoning in children (Fazio et al. 2014; Ruggeri et al. 2018), the current series of experiments marks the first systematic attempt to uncover the potential links between the analog representations of numerical magnitude and adult's ability to approximate probabilities. This is important because adults have a more sophisticated understanding of the proportional nature of probability as well as more precise mental representations of numerical magnitude. We found that participants with better number approximation abilities were also more successful at judging non-symbolic proportions. Although this finding is correlational, we also find that the same non-numerical stimulus features influencing perceived numerosity also influence perceived probability. In addition, we found that non-numerical features known to influence perceived numerosity, namely object size and spacing, also influence judgements about probability. Analyses of model coefficients from the psychophysical model indicated that although participants relied mostly on proportion to make their decisions, the model coefficients for size and spacing were significantly greater than 0 suggesting that these features influenced perceived probability. Together, these findings suggest that more precise representations of number can improve the precision of probability approximations.

Our results also shed light on the ubiquitous nature of heuristic decision rules in human probability judgments. In all four experiments we find that although participants' responses are for the most part accurate, they are sometimes influenced by simple heuristics which allow for more rapid yet less accurate decision-making. The use of heuristics and biases in human decision-making is pervasive and the literature on this topic is vast. While a great deal of classic research has focused on the use of heuristics when reasoning about conditional probability (Kahneman & Tversky, 1972; Tvesrky & Kahneman, 1973; Tversky & Kahneman, 1974) more recent work has shown that adults also use heuristic decision rules when presented with simple probability judgments (Alonso & Fernández-Berrocal, 2003; Denes-Raj & Epstein, 1994; Kirkpatrick & Epstein, 1992; Pacini & Epstein, 1999; O'Grady et al. 2016). Developmental research suggests that by the time children are 8 years old, they begin to understand the proportional nature of simple probability judgments involving small, countable sets of objects and by the time they are 10 years old the majority of children no longer use heuristic decision rules.(Falk et al. 2012). Furthermore, when adults are presented with the same simplified probability problems involving small numbers of countable sets they do not demonstrate a heuristic bias (Chapman, 1975). Importantly, much of the research reporting adults' use of heuristic decision rules involved tasks in which adult participants were not provided with exact numerical information (Alonso-Diaz, Piantadosi, Hayden, & Cantlon, 2018; Eckert et al., 2018; Fazio et al., 2014; O'Grady et al. 2016) or used probabilities that were difficult to discriminate because they were close together or equal in magnitude (Alonso & Fernández-Berrocal, 2003; Denes-Raj & Epstein, 1994; Kirkpatrick & Epstein, 1992; Pacini & Epstein, 1999) suggesting that adults are capable of accurate probability judgments when they have complete access to numerical information. In the current series of studies we report results which converge with these previous findings: when probabilities are difficult to discriminate, people use a heuristic strategy by choosing the group with the greatest number of target marbles.

Why are adults capable of making accurate probability judgments when presented with simple computations yet they show heuristic biases when presented with large groups of marbles presented for a short amount of time? New theoretical approaches to cognitive modeling have called for resource-rational analyses of human cognition (Griffiths, Lieder, & Goodman, 2015; Lieder & Griffiths, 2019). From this perspective, non-optimal decision making results from the rational use of their limited cognitive resources given environmental constraints. Simply put, when people are presented with a difficult task with several possible solutions, they select a strategy that is reasonably accurate given their finite cognitive resources (Lieder & Griffiths, 2019). In accordance with this perspective, recent research on strategy selection suggests that children and adults are keenly aware of time-accuracy tradeoffs in problem-solving tasks and will switch strategies in accordance with rational principles regarding available cognitive resources (Lieder & Griffiths, 2017). With respect to the current series of experiments, a resource-rational approach would argue that adults understand the proportional nature of probability but when their ability to compute probability is constrained by time (i.e. 750 – 1500 ms), and limited cognitive resources (i.e. inexact, analog magnitude representations of number) they fall back on simple magnitude information such as area and absolute number.

Although our results offer novel insights into the human decision-making process with regards to perceived probability, there are several limitations of the current experimental design. First, the current series of experiments involved purely non-symbolic forms of probability yet much of the probabilities humans routinely engage with involve a mix of both symbolic and non-symbolic formats. Recent research has shown that non-symbolic rational number processing

abilities predicts performance on symbolic mathematics measures (Matthews, Lewis & Hubbard, 2016). Although previous research has investigated rational number processing involving a mix of both symbolic fractions and non-symbolic ratios (Matthews & Chesney, 2015), we are not aware of any experiments that have incorporated symbolic and non-symbolic representations of probability. In order to enhance the ecological validity of the current series of experiments, future work may present participants with a combination of both symbolic and non-symbolic representations of probability. Such a manipulation may also shed further light on the role of number approximation in probability estimation: Do people give more weight to symbolic representations of number probability as opposed to approximate representations when computing probabilities? Second, the current series of studies required participants to make a decision based on a single random draw based on proportion, yet a great deal of probabilistic data encountered in the real world involved sequential probabilities. Future work will investigate the psychophysical properties of decisions made during both sequential and single event probabilities. A third potential limitation of the current study design relates to learning in probabilistic contexts. Although some developmental research has studied the influence of feedback on children's choices in 2-alternative forced-choice random draw tasks, we are not aware of any studies investigating the role of feedback on adults' use of heuristic decision rules in these tasks. What influence does feedback have on adults' decisions and can they be trained to use more accurate strategies in probabilistic reasoning tasks?

Results of the current series of experiments converge with previous research demonstrating the ratio dependence of probability estimation based proportions (Eckert et al., 2018; Fazio et al., 2014; Alonso-Diaz, Piantadosi, Hayden, & Cantlon, 2018; O'Grady & Xu, 2019). We extend these findings by providing evidence that human adults utilize number approximation and heuristic decision rules when making decisions based on probability. Future work will investigate a larger range of stimulus presentations times as well as the factors that influence the relative weighting of numerical magnitude information and heuristics in order to shed further light on human decision-making in probabilistic contexts.

2.7 Appendix

2.7.1 Additional information for Experiment 1

This section contains additional information about the methods and results reported in Experiment 1.

Table 2.A1 Full description of the contents of each image used in Experiment 2.1

	Correct Choice				Incorrect Choice				
Trial Type	Number of target marbles	Number of Non-target marbles	Total number of marbles	Probability	Number of target marbles	Number of Non-target marbles	Total number of marbles	Probability	Ration of proportions
Total equal	55	45	100	0.55	50	50	100	0.5	1.1
Target equal	50	40	90	0.5556	49	49	98	0.5	1.111
Target equal	59	25	84	0.7024	59	39	98	0.602	1.167
Total equal	70	30	100	0.7	60	40	100	0.6	1.167
Foil	70	30	100	0.7	42	28	70	0.6	1.167
Total equal	55	45	100	0.55	45	55	100	0.45	1.222
Target equal	45	37	81	0.5556	45	54	99	0.455	1.221
Target equal	55	14	69	0.7971	55	37	92	0.598	1.333
Total equal	80	20	100	0.8	60	40	100	0.6	1.333
Foil	80	20	100	0.8	27	18	45	0.6	1.333
Foil	75	25	100	0.75	28	22	50	0.56	1.339
Target equal	53	13	66	0.803	53	43	96	0.552	1.455
Total equal	80	20	100	0.8	55	45	100	0.55	1.455
Total equal	60	40	100	0.6	40	60	100	0.4	1.5
Target equal	40	26	66	0.6061	40	59	99	0.404	1.5
Foil	55	44	99	0.5556	19	35	54	0.352	1.578
Target equal	20	8	28	0.7143	20	28	48	0.417	1.713
Target equal	39	17	56	0.6964	39	59	98	0.398	1.75
Total equal	70	30	100	0.7	40	60	100	0.4	1.75
Total equal	55	45	100	0.55	30	70	100	0.3	1.833
Target equal	30	24	54	0.5556	30	69	99	0.303	1.834
Foil	69	29	98	0.7041	15	27	42	0.357	1.972
Target equal	30	20	50	0.6	30	70	100	0.3	2
Total equal	60	40	100	0.6	30	70	100	0.3	2
Total equal	70	30	100	0.7	30	70	100	0.3	2.333
Target equal	30	12	42	0.7143	29	69	98	0.296	2.413
Total equal	80	20	100	0.8	30	70	100	0.3	2.667
Target equal	29	7	36	0.8056	29	67	96	0.302	2.668
Target equal	24	8	32	0.75	24	72	96	0.25	3
Total equal	75	25	100	0.75	25	75	100	0.25	3
Foil	72	24	96	0.75	7	21	28	0.25	3
Total equal	70	30	100	0.7	20	80	100	0.2	3.5
Target equal	20	5	25	0.8	20	80	100	0.2	4
Total equal	80	20	100	0.8	20	80	100	0.2	4
Foil	80	20	100	0.8	4	16	20	0.2	4
Foil	77	19	96	0.8021	2	10	12	0.167	4.803
Foil	50	50	100	0.5	1	9	10	0.1	5
Foil	69	29	98	0.7041	1	7	8	0.125	5.633
Total equal	90	10	100	0.9	15	85	100	0.15	6
Target equal	15	2	15	1	14	76	90	0.156	6.41
Total equal	80	20	100	0.8	10	90	100	0.1	8
Target equal	10	2	12	0.8333	10	86	96	0.104	8.013
Target equal	9	1	10	0.9	9	81	90	0.1	9
Total equal	90	10	100	0.9	10	90	100	0.1	9
Target equal	5	5	10	0.5	5	95	100	0.05	10
Total equal	50	50	100	0.5	5	95	100	0.05	10
Total equal	55	45	100	0.55	5	95	100	0.05	11
Target equal	5	4	9	0.5556	5	94	99	0.051	10.894
Total equal	70	30	100	0.7	5	95	100	0.05	14
Target equal	5	2	7	0.7143	5	93	98	0.051	14.006

Reaction time analyses for Experiment 1. Comparisons of linear models for reaction time revealed that the model predicting reaction time from ratio of proportions, trial type and the interaction of ratio of proportions and trial type ($AIC_{Interaction}$ =23,494.57) outperformed the null model (AIC_{Null} =23,583.81; χ^2= 95.25; *df* = 3; $p < .001$), the models predicting reaction time from ratio of proportions alone (AIC_{RP} =23,585.22; χ^2= 94.65; *df* = 2; $p < .001$) and trial type alone (AIC_{TT} =23,501.19; χ^2= 10.63; *df* = 2; $p < .001$) as well as the model predicting reaction time from ratios of proportions and trial type without an interaction (AIC_{RP+TT} =23,503.04; χ^2= 10.47; *df* = 1; $p < .01$).

Inspection of the model coefficients ($\beta_{Intercept}$ = 537.26; SE = 36.08; 95% CI [466.54, 607.97]). revealed that increasing ratio of proportions led to small increases in reaction time for *target equal* trials (β_{RP} = 13.31; SE = 4.48; 95% CI [4.52, 22.10]) and a decrease for *total equal* trials (β_{RP*TT} = -8.84; SE = 2.73; 95% CI [-14.19, -3.49]]). Holding ratio of proportions constant, *total equal* trials (β_{TT} = -54.52; SE = 14.93; 95% CI [-83.77, -25.26]) showed lower reaction times compared to *target equal* trials.

2.7.2 Additional information for Experiment 2

The following sections provide additional information about the methods and results reported for Experiment 2.

Table 2.A2 Full description of the contents of each image used in Experiment 2.2

	Correct Choice				Incorrect Choice				
Trial Type	Number of target marbles	Number of Non-target marbles	Total number of marbles	Probability	Number of target marbles	Number of Non-target marbles	Total number of marbles	Probability	Ration of proportions
Area anti-correlated	11	9	20	0.55	10	10	20	0.5	1.1
Number vs proportion	11	9	20	0.55	19	19	38	0.5	1.1
Total equal	11	9	20	0.55	10	10	20	0.5	1.1
Area anti-correlated	10	10	20	0.5	9	11	20	0.45	1.11
Number vs proportion	10	10	20	0.5	18	22	40	0.45	1.11
Total equal	10	10	20	0.5	9	11	20	0.45	1.11
Area anti-correlated	18	22	40	0.45	16	24	40	0.4	1.13
Number vs proportion	18	22	40	0.45	24	36	60	0.4	1.13
Total equal	18	22	40	0.45	16	24	40	0.4	1.13
Area anti-correlated	11	9	20	0.55	9	11	20	0.45	1.22
Number vs proportion	11	9	20	0.55	18	22	40	0.45	1.22
Total equal	11	9	20	0.55	9	11	20	0.45	1.22
Area anti-correlated	12	3	15	0.8	9	6	15	0.6	1.33
Number vs proportion	12	3	15	0.8	18	12	30	0.6	1.33
Total equal	12	3	15	0.8	9	6	15	0.6	1.33
Area anti-correlated	16	4	20	0.8	11	9	20	0.55	1.45
Number vs proportion	16	4	20	0.8	22	18	40	0.55	1.45
Total equal	16	4	20	0.8	11	9	20	0.55	1.45
Area anti-correlated	12	8	20	0.6	8	12	20	0.4	1.5
Number vs proportion	12	8	20	0.6	18	27	45	0.4	1.5
Total equal	12	8	20	0.6	8	12	20	0.4	1.5
Area anti-correlated	12	4	16	0.75	8	8	16	0.5	1.5
Number vs proportion	12	4	16	0.75	20	20	40	0.5	1.5
Total equal	12	4	16	0.75	8	8	16	0.5	1.5
Area anti-correlated	19	1	20	0.95	11	9	20	0.55	1.73
Number vs proportion	19	1	20	0.95	11	9	20	0.55	1.73
Total equal	20	1	21	0.95	33	27	60	0.55	1.73
Area anti-correlated	19	1	20	0.95	10	10	20	0.5	1.9
Number vs proportion	19	1	20	0.95	27	27	54	0.5	1.9
Total equal	19	1	20	0.95	10	10	20	0.5	1.9
Area anti-correlated	10	10	20	0.5	5	15	20	0.25	2
Number vs proportion	10	10	20	0.5	18	54	72	0.25	2
Total equal	10	10	20	0.5	5	15	20	0.25	2
Area anti-correlated	8	12	20	0.4	3	17	20	0.15	2.67
Number vs proportion	8	12	20	0.4	15	85	100	0.15	2.67
Total equal	8	12	20	0.4	3	17	20	0.15	2.67
Area anti-correlated	12	8	20	0.6	4	16	20	0.2	3
Number vs proportion	12	8	20	0.6	17	68	85	0.2	3
Total equal	12	8	20	0.6	4	16	20	0.2	3
Area anti-correlated	10	10	20	0.5	3	17	20	0.15	3.33
Number vs proportion	10	10	20	0.5	15	85	100	0.15	3.33
Total equal	10	10	20	0.5	3	17	20	0.15	3.33
Area anti-correlated	12	3	15	0.8	3	12	15	0.2	4
Number vs proportion	12	3	15	0.8	17	68	85	0.2	4
Total equal	12	3	15	0.8	3	12	15	0.2	4
Area anti-correlated	4	6	10	0.4	1	9	10	0.1	4
Number vs proportion	4	6	10	0.4	9	81	90	0.1	4
Total equal	4	6	10	0.4	1	9	10	0.1	4
Area anti-correlated	15	5	20	0.75	3	17	20	0.15	5
Number vs proportion	6	2	8	0.75	12	68	80	0.15	5
Total equal	15	5	20	0.75	3	17	20	0.15	5
Area anti-correlated	19	1	20	0.95	24	135	159	0.15	6.33
Number vs proportion	19	1	20	0.95	3	17	20	0.15	6.33
Total equal	19	1	20	0.95	3	17	20	0.15	6.33
Area anti-correlated	8	2	10	0.8	1	9	10	0.1	8
Number vs proportion	4	1	5	0.8	9	81	90	0.1	8
Total equal	16	4	20	0.8	2	18	20	0.1	8
Area anti-correlated	11	2	13	0.85	16	144	160	0.1	8.5
Number vs proportion	17	3	20	0.85	2	18	20	0.1	8.5
Total equal	17	3	20	0.85	2	18	20	0.1	8.5
Area anti-correlated	9	1	10	0.9	1	9	10	0.1	9
Number vs proportion	9	1	10	0.9	14	126	140	0.1	9
Total equal	18	2	20	0.9	2	18	20	0.1	9
Area anti-correlated	18	1	19	0.95	20	180	200	0.1	9.5
Number vs proportion	19	1	20	0.95	2	18	20	0.1	9.5
Total equal	19	1	20	0.95	2	18	20	0.1	9.5

Reaction time analyses for Experiment 2. Results of model comparisons for reaction time data revealed that the model predicting reaction time from ratio of proportions and trial type without the interaction of ratio of proportions and trial type (AIC_{RP+TT} =134,992.82) outperformed the null model (AIC_{Null} =135,129.10; χ^2= 142.28; *df* = 3; $p < .001$), the models predicting reaction time from ratio of proportions alone (AIC_{RP} =135,054.27; χ^2= 65.46; *df* = 2; $p < .001$) and trial type alone (AIC_{TT} =135,068.97; χ^2= 78.16; *df* = 1; $p < .001$). The model for main effects did not differ significantly from the model with interactions (AIC_{RP*TT} =135,014.84; χ^2= 0; *df* = 2; p = 1) which indicates that it is the preferred model as it explains the same amount of variance with fewer parameters.

Inspection of the model coefficients ($\beta_{Intercept}$ = 424.74; SE = 28.38; 95% CI [369.11, 480.38]) revealed that an increase in ratio of proportions led to small decreases in reaction time (β_{RP} = -6.89; SE = 0.78; 95% CI [-8.41, -5.36]) and that participants were about 40 ms slower on both *number vs proportion* trials (β_{NvP} = 37.26; SE = 5.29; 95% CI [26.89, 47.62]) and *area-anticorrelated* trials (β_{AA} = 37.04; SE = 5.30; 95% CI [26.64, 47.43]) compared to *total equal* trials.

2.7.3 Additional information for Experiment 3

The following sections provide additional information about the methods and results reported for Experiment 3.

Reaction time analyses for Experiment 3. As with Experiments 1 & 2 model comparisons for reaction time data revealed that the model predicting reaction time from ratio of proportions and trial type without the interaction of ratio of proportions and trial type (AIC_{RP+TT} =107,489.09) outperformed the null model (AIC_{Null} =107,514.19; χ^2= 31.10; *df* = 3; $p < .001$), the models predicting reaction time from ratio of proportions alone (AIC_{RP} =107,495.57; χ^2= 10.48; *df* = 2; $p < .001$) and trial type alone (AIC_{TT} =107,507.78; χ^2= 20.68; *df* = 1; $p < .001$). The model for main effects did not differ significantly from the model with interactions (AIC_{RP*TT} =107,491.10; χ^2= 1.99; *df* = 2; p = 0.37).

Inspection of the model coefficients ($\beta_{Intercept}$ = 442.15; SE = 34.83; 95% CI [373.87, 510.43]) revealed that an increase in ratio of proportions led to small decreases in reaction time (β_{RP} = -3.88; SE = 0.85; 95% CI [-5.56, -2.21]) and that participants were about 15 ms slower on both *number vs proportion* trials (β_{NvP} = 14.16; SE = 5.78; 95% CI [2.83, 25.50]) and *area-anticorrelated* trials (β_{AA} = 17.84; SE = 5.83; 95% CI [6.40, 29.27]) compared to *total equal* trials.

2.7.4 Additional information for Experiment 4

The following sections provide additional information about the methods and results reported for Experiment 4.

Reaction time analyses for Experiment 4. Model comparisons for reaction time data in Experiment 4 revealed that the model predicting reaction time from ratio of proportions and trial type without the interaction of ratio of proportions and trial type (AIC_{RP+TT} =120,554.12) outperformed the null model (AIC_{Null} =120,617.45; χ^2= 69.32; *df* = 3; $p < .001$), the models

predicting reaction time from ratio of proportions alone (AIC_{RP} =120,561.67; χ^2= 11.55; *df* = 2; $p < .001$) and trial type alone (AIC_{TT} =120,610.10; χ^2= 57.97; *df* = 1; $p < .001$). The model for main effects did not differ significantly from the model with interactions (AIC_{RP*TT} =120,554.09; χ^2= 4.03; *df* = 2; $p = 0.13$) which indicates that it is the preferred model as it explains the same amount of variance with fewer parameters.

Inspection of the model coefficients ($\beta_{Intercept}$ = 406.89; SE = 22.88; 95% CI [362.06, 451.73]) revealed that an increase in ratio of proportions led to small decreases in reaction time (β_{RP} = -5.23; SE = 0.69; 95% CI [-6.57, -3.89]) and that participants were about 8 ms slower on *number vs proportion* trials (β_{NvP} = 8.05; SE = 4.68; 95% CI [-1.12, 17.21]) and 16 ms slower on *area-anticorrelated* trials (β_{AA} = 15.95; SE = 4.69; 95% CI [6.75, 25.15]) compared to *total equal* trials.

Chapter 3

The Development of Non-symbolic Probability Judgments in Children

3.1 Introduction

Children experience a great deal of probabilistic data in everyday life, and both developmental psychologists and educators have found probabilistic reasoning to be a fertile domain for understanding the development of numerical cognition. Throughout development, children encounter a wealth of numerical and non-numerical data and must integrate these data to make rapid judgments, often based on limited information. Understanding how children leverage their intuitive understanding of number and probability to make decisions in a complex world can provide insights that are relevant to a broad range of fields from perception and decision making to formal mathematics education.

Probabilistic reasoning refers to a broad range of abilities related to uncertainty such as understanding randomness, appropriately analyzing sample spaces, reasoning about correlation, and formally quantifying probability (see Bryant and Nunes, 2012, for a thorough and insightful discussion). This is a broad literature with numerous unanswered research questions. In this paper, we focus primarily on children's estimation of the probability of discrete events and we aim to chart the developmental trajectory of these abilities. We begin by briefly reviewing the relevant literatures on the development of probabilistic and proportional reasoning abilities as well as the approximate number system. We then present the results of two experiments designed to investigate the influence of numerical and non-numerical stimulus features on children's probability judgments and to track the development of the ability to reason about probability based on proportion.

For discrete outcomes, probability is computed as a proportion of target outcomes to all possible outcomes. While a ratio formally describes a relation between two quantities, a proportion is used to assess the equality of two ratios. Although both proportions and ratios can be used to compare probabilities of binary outcomes, comparing ratios can be difficult when the total number of possible outcomes differs between two options. For example, imagine a child is presented with two groups of red and white marbles and asked to choose the group that is most likely to yield a red marble from a single random draw. Imagine further that one group has 13 red marbles and 7 white marbles while the second group has 15 red marbles and 9 white marbles. Representing these choices as ratios provides the observer with part:part comparisons, (13:7) and (15:9). While adults may be adept at computing odds based on ratios, children have difficulty performing these computations and often make their choice based on the group with the larger number of target items (in this case, red marbles) rather than on the proportion of red marbles to total marbles. Proportions facilitate this comparison by formalizing the relation between the parts (subsets of outcomes) and the whole (all possible outcomes). In our example, the two proportions would be 13/20 (0.65) and 15/24 (0.625). The first group has a slightly higher chance of yielding

a red marble on a single random draw. Thus, the ability to compute proportions can help children accurately judge the equivalence of two probabilities.

For decades, many studies have used the 2-alternative forced-choice (2AFC) random draw task to investigate children's predictions about single and sequential random draws (Falk, Falk, & Levin, 1980; Falk, Yudilevich-Assouline, & Elstein, 2012; Piaget & Inhelder, 1975; Siegler, Strauss, & Levin, 1981; Yost, Siegel, & Andrews, 1962). In this task, children are typically presented with two groups of multiple objects and asked to select the group with the best chance of getting a preferred object. Recently, Falk et al.(2012) conducted a comprehensive study of probabilistic decision making strategies using the 2AFC random draw task with 6- to 12-year-old children. Findings from this study revealed that young children often choose the group with the greatest number of target objects regardless of the total number of objects until around 8 years of age when children begin to attend to the whole set of possible outcomes rather than simply the number of target outcomes. These findings indicate that younger children have difficulty reasoning about probability based on proportion: instead of relating a part (a subset of outcomes) to the whole (all possible outcomes) for each choice and choosing the group with the larger proportion of target outcomes, children merely compare the number of target outcomes in each choice and choose the group with more target outcomes.

Much like the research on probabilistic reasoning, research on proportional reasoning has also shown that children's ability to reason about proportion greatly improves over the school-age years (Spinillo & Bryant, 1999; Mix, Levine & Huttenlocher 1999; Mohring, Newcombe, Levine, & Fricke, 2015; Singer-Freeman & Goswami, 2001). Many of these studies have deployed the proportional match to sample task in which a child is first presented with a proportional 'target' stimuli then presented with several similar proportions from which they should select the item that matches the proportions of the target stimulus. These methods often compare children's choices when they are presented with proportions in a discrete format (i.e. discrete units of juice and water) to their choices when presented with the same proportions in a continuous format (i.e. portions of juice and water that do not have discrete units). Using this method, researchers have reported a common error in which children choose the item based on matching parts rather than matching proportions (Boyer et al., 2008; Boyer & Levine, 2012;). This error is similar to the types of incorrect choices made by children in probabilistic reasoning studies discussed earlier (Piaget & Inhelder, 1975; Falk et al. 2012). In the proportional reasoning literature, research has shown that these errors are most often observed when children are presented with stimuli containing discrete, countable parts (Boyer et al., 2008; Boyer & Levine, 2012; Boyer & Levine, 2015; Boyer, Levine & Huttenlocher, 2008; Hurst & Cordes, 2018; Jeong, Levine, & Huttenlocher, 2007). These findings suggest that young children's proportional judgments are influenced by their knowledge of whole numbers (Mix, Levine, & Huttenlocher, 1999; Sohpian & Wood, 1997; Sophian, 2000). Based on the findings from the proportional reasoning literature, children can make accurate proportional matches when presented with proportions in a continuous format but they show a bias toward comparisons of parts when they are presented with proportions in a discrete format. In the current paper we seek to chart the developmental trajectory of the ability to compute the probability of discrete events based on proportion.

Humans have remarkable abilities for reasoning about numerical magnitude (Feigenson, Dehaene, and Spelke, 2004; Dehaene, 1997 / 2011). Our ability to rapidly form accurate approximations of numerical magnitude is often referred to as the approximate number system (ANS) and can be observed within the first year of life (Dehaene et al., 1998; Izard, Sann,

Spelke, & Streri, 2009; Lipton & Spelke, 2003; Wood & Spelke, 2003; Xu & Spelke, 2000; Xu 2003; Xu, Spelke, & Goddard, 2005). In addition to number discrimination, infants form expectations about approximate addition and subtraction (Chiang & Wynn, 2000; McCrink & Wynn, 2004) and can even discriminate ratios (McCrink & Wynn, 2007). Furthermore, young children's performance in non-symbolic multiplication and division tasks (McCrink & Spelke, 2010, 2016) suggests that ANS representations play a role in arithmetical reasoning even when children have not been formally trained to use algorithms for symbolic multiplication and division.

Decades of research on numerical processing has shown that both humans and non-human animals are capable of forming abstract representations of number (Dehaene, Dehaene-Lambertz, & Cohen, 1998; Moyer & Landauer, 1967; Pica, Lemer, Izard, & Dehaene, 2004; Whalen, Gallistel, & Gelman, 1999). The rapid and inexact nature of ANS representations follows Weber's Law (Halberda & Feigenson, 2008; Pica et al., 2004; Whalen et al., 1999) and thus demonstrates ratio dependence: the ability to discriminate two sets of objects based on number depends upon the ratio of the magnitudes of those sets. As a result, an observer's ability to discriminate sets of objects based on numerical magnitude depends on the distance between the two numbers along a mental number line.

The acuity of an individual's ANS representations can be measured using a psychophysical design in which sets of colored dots (i.e. yellow dots vs blue dots) are presented to an observer who is asked to identify the set that contains the largest number of dots. Importantly, experimenters using this method manipulate the ratios of the two sets of dots and the 'distance effect' is observed when smaller ratios are more difficult to discriminate than larger ratios.

The goal of the current series of experiments is to chart the developmental trajectory of probabilistic reasoning by measuring the acuity of children's ability to discriminate probabilities based on proportions. We investigate the possibility that children's judgments about probability based on proportion will demonstrate ratio dependence similar to results reported in the ANS literature as well as whether their probability judgments are influenced by the same erroneous, part:part reasoning reported in the proportional reasoning literature.

In adults, ANS acuity is correlated with performance on approximate probability judgment tasks (O'Grady, Starr, Griffiths, & Xu, submitted). Researchers have reported distance effects in ratio magnitude comparison tasks framed as probability judgments for 12-year-old children using methods adapted from the psychophysics of number perception (Fazio, Bailey, Thompson, & Siegler, 2014) and developmental researchers have also reported distance effects in younger children using a sequential probability task (Boyer, 2007). The current study marks the first attempt to trace the developmental trajectory of children's probability approximation abilities.

Previous research on children's probabilistic reasoning has employed tasks in which children are presented with small number of countable sets of objects which may have primed them to focus on the absolute number of target objects rather than the relative frequency of target objects (e.g., Falk et al. 2012). Based on the findings from the proportional reasoning literature we hypothesize that young children are capable of making rapid and accurate approximations of probability based on proportions. From this hypothesis, we make three predictions: First, children's probabilistic discrimination abilities will demonstrate ratio dependence (i.e. as two proportions move further apart on the mental number line, they will become easier to discriminate; also known as the distance effect). Second, the ability to discriminate probabilities

will improve with age since both proportional reasoning and number approximation acuity improves with age. Third, based on previous research on probability judgments and proportional reasoning, we predict that probability discrimination will be influenced by the same non-numerical features known to influence perceived numerosity (e.g., size of the objects). Furthermore, we investigate whether children's non-symbolic probability estimates will be influenced by the same heuristic decision rules reported in previous research that required counting (Falk et al., 2012); that is, whether children sometimes use only the number of target objects when estimating probability as opposed to the correct proportion strategy.

3.2 Experiment 1

3.2.1 Methods

Participants. Sixty 6- to 7-year-old children were recruited from local public schools and museums in the San Francisco Bay area. According to the National Center for Educational Statistics, (NCES, 2018), the schools in which we conducted the current series of experiments serve children from a range of racial and ethnic backgrounds (School A: 14% Asian, 5% Black, 11% Latinx, 1% Native Hawaiian, 57% White & 12% Mixed race/ethnicity; School B: 28% Asian, 8% Black, 16% Latinx, 38% White & 10% Mixed race/ethnicity). Although we did not collect data on socioeconomic status, we conduct our experiments at local museums on free admission days in order to recruit families from a range of socioeconomic backgrounds. According to data from the United States Census Bureau, the median incomes of the three communities in which data were collected are $70,393 per year, $92,670 per year, and $140,640 per year (U.S. Census Bureau, 2018) indicating that children in the current sample came from middle to upper-middle class households.

A total of twelve children (10 six-year-olds and 2-seven-year-olds) were excluded because they did not pass practice trials meant to ensure that participants understood the task. The remaining sample of participants consisted of 24 six-year-olds (N = 24; Mean age = 6.28; SD =0.30; 19 female) and 24 seven-year-olds (N = 24; Mean age = 7.62; SD =0.36; 20 female). Target sample size (N=48) was determined based on previous research with similar tasks (Fazio et al., 2014; Halberda et al. 2008) as well as the additional constraint of ensuring that only children who passed practice trials were included in the final sample. Since Halberda et al. (2008) investigated age related differences in a simple dot approximation task with 16 children in each of five age groups and Fazio et al. (2014) collected data for a sample of 53 twelve-year-olds in a ratio comparison task but did not seek to investigate age related differences, we decided to split the difference and test two age groups with 24 children each.

Material. The images for the task were rendered using Blender 2.72, 3D animation software (http://www.blender.org/). Each image contained two groups of red and white marbles divided by a black partition. Since the goal of this experiment was to investigated the psychophysical properties of probability judgments, we created images with a wide range of proportions. The ratio of the proportions presented in each image ranged from 1.1 (55% vs. 50%) and 14 (70% vs. 5%). Table 1 contains the ratios of the proportions used in Experiment 1. For each ratio of proportion, two trial types were created. The *total equal* trials contained the same total number of marbles on each side of the partition, while the *target equal* trials contained the same number of target color marbles on each side with the ‘losing’ group containing more non-target marbles. In

total, there were 100 marbles in each group in the *total equal* trials while *target equal* trials contained one group with 100 marbles matching the 'losing' proportion and another group containing an equal number of target color marbles and enough non-target color marbles to match the 'winning' proportion. Importantly, the difference in the total amount of marbles created a large contrast between the field areas of the 'winning' and 'losing' groups. In order to reduce the chances that this contrast could cue participants to choose the group with the smallest field area we also created an additional set of 'foil' trials in which the 'winning' group contained more marbles and thus a larger field area than the 'losing' group. Figure 1 contains a visual schematic of the procedure with an example of each trial types as well as an example foil trial image.

Table 3.1 Proportion presented in each trial of Experiment 1.

Proportion group 1	Proportion group 2	Ratio of Proportions
0.55	0.50	1.10
0.70	0.60	1.17
0.55	0.45	1.22
0.80	0.60	1.33
0.80	0.55	1.45
0.60	0.40	1.50
0.70	0.40	1.75
0.55	0.30	1.83
0.60	0.30	2.00
0.70	0.30	2.33
0.80	0.30	2.67
0.75	0.25	3.00
0.70	0.20	3.50
0.80	0.20	4.00
0.90	0.15	6.00
0.80	0.10	8.00
0.90	0.10	9.00
0.50	0.05	10.00
0.55	0.05	11.00
0.70	0.05	14.00

Note. Ratios of Proportions are rounded to 2 decimal points.

Procedure. After their parents signed a written consent form approved by the University of California Berkeley Committee for the Protection of Human Subjects (CPHS) children were asked to provide verbal assent to participate in the study. Children were then seated in front of a

MacBook Pro laptop (OSX; Screen resolution 1280 x 800) and were told that they were going to play a game in which they would help Big Bird collect marbles. Half of the children were instructed to collect red marbles and the other half were asked to collect white marbles.

An experimenter explained that Big Bird could not see the contents of the bags of marbles and that he would take a single marble randomly from the bag that the child chooses. The child was then reminded that Big Bird preferred either red or white marbles and that they should choose the group which was best for getting a marble of that color. The experimenter then told the children that one choice was always better than the other and that some of the trials might seem easy while others may be more difficult. Furthermore, if they were uncertain about which group to choose, they should try to make their best guess. In order to reduce the influence of age related differences in formal understanding of the words 'probability' and 'proportion', the experimenter never explicitly mentioned the words 'probability' or 'proportion' during the instructions. Children were then presented with four practice trials with two groups of marbles, one containing all red marbles while the other contained all white marbles. Participants were told that the practice trials were intentionally easy and were meant to teach them how the game worked.

Each participant was presented with 40 test trials and 10 foil trials in one of two semi-randomized orders using the psychophysics toolbox (Brainard, 1997,Pelli (1997); Kleiner, Brainard, & Pelli, 2007) written for the MatLab programing language. Since previous research using a similar design presented images with fewer objects for 1320ms (Fazio et al., 2014) we decided to present the images for 1500ms to allow our younger participants more time. Following stimulus presentation, participants saw a screen containing the Big Bird character flanked by two bags labeled with a blue '1' and a green '2'. Participants were instructed to press a key marked by a sticker matching the color of the number on the bag they wanted Big Bird to draw a marble from. Intermission videos in the form of a 30 second animation were used to give children a break during the game and were presented after the 15th, 30th, and 40th trials. Importantly, children did not receive feedback on their choices until the end of the game at which point every child saw the same screen containing 40 white or red marbles and was told 'Wow, you did really good! Look how many red/white marbles you got!'. The computer collected both reaction time and participant choice for each trial. Once the participant completed the last trial, they saw a screen containing 40 marbles that matched their target color and were told that these were the marbles that they had collected during the game. A visual schematic of the procedure is presented in Figure 3.1.

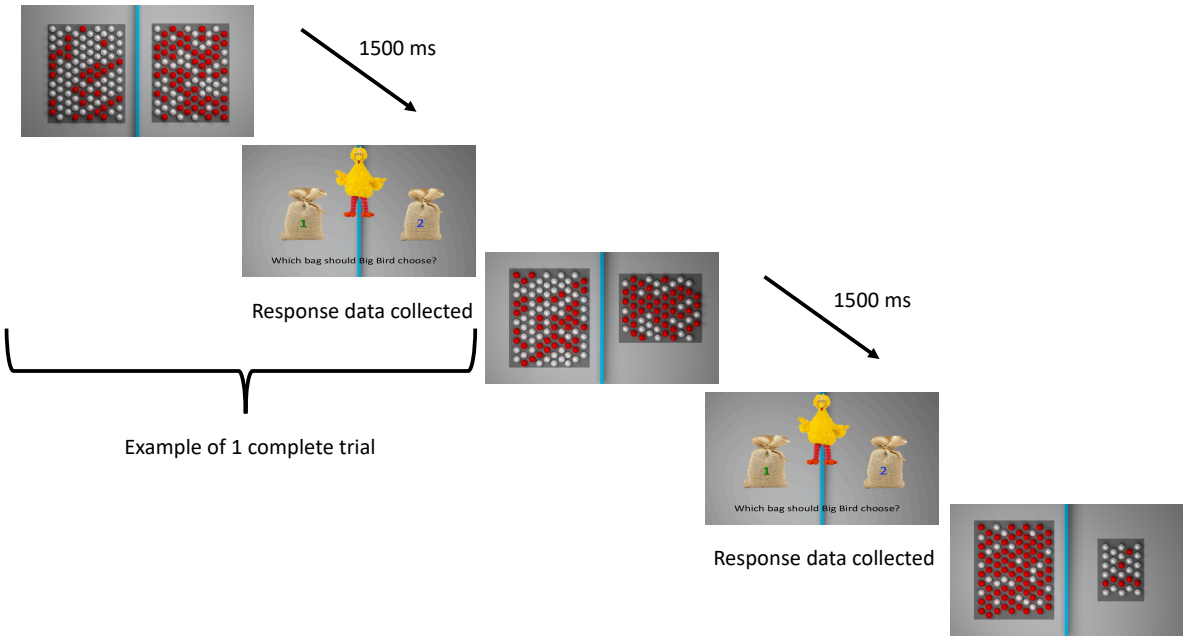

Figure 3.1 A visual schematic of the experimental procedure used in experiment 1. The sample image at the top presents a total equal trial, the third image from the top presents a sample target equal trial, and the image at the very bottom of the figure presents a sample of the foil trials used to prevent participants from learning to choose the group with the smaller amount of marbles.

3.2.2 Results

Using the binomial exact test we find that performance on foil trials was above chance for both 6-year-olds ($probability\ of\ success = 0.79$, 95% CI [0.73, 0.84], $p < .001$) and 7-year-olds ($probability\ of\ success = 0.97$, 95% CI [0.94, 0.99], $p < .001$) suggesting that children did not learn to merely select the smallest group when presented with groups of different sizes. Foil trials were excluded from the remainder of analyses.

Reaction time. Reaction time data were cleaned for outliers by excluding reaction times which were either greater or less than 3 Median Absolute Deviations (MADs) from each participant's median reaction time. Since the median is relatively insensitive to the effects of outliers compared to the mean, this method is thought to be a superior method for identifying outlying reaction time data (Leys, Ley, Klein, Bernard, & Licata, 2013). Use of this procedure resulted in the exclusion of 198 of the 1920 total trials (10.31%). In order to report the most accurate representations of the data, all of analyses reported in this paper were conducted on the dataset in which trials in which outlying reaction time were excluded. Exclusion of these data do not change the results for any of the following analyses including general accuracy and statistical modeling. Results of the same analyses conducted on the complete dataset are reported in the Online Supporting Information. Comparisons of performance for all included trials revealed that the reaction time for both age groups was significantly faster on the *total equal* trials (6-year-olds: M = 1,121.67 ms, SD = 923.22 ms; 7-year-olds: M =642.57 ms, SD = 508.97) compared to the *target equal* trials (6-year-olds: M = 1,366.98ms, SD = 1,163.47; $\Delta M = 245.31$, 95% CI

$[-386.98, -103.63]$, $t(781.47) = -3.40$, $p = .001$; 7-year-olds: M = 758.51 ms, SD = 547.90; $\Delta M = 115.94$, 95% CI $[-186.55, -45.33]$, $t(852.71) = -3.22$, $p = .001$).

General accuracy. Children in both age groups performed significantly above chance on both *total equal* (6-year-olds: $probability\ of\ success = 0.78$, 95% CI [0.74, 0.82], $p < .001$; 7-year-olds: $probability\ of\ success = 0.91$, 95% CI [0.88, 0.93], $p < .001$; binomial exact test) and *target equal* trial types (6-year-olds: $probability\ of\ success = 0.67$, 95% CI [0.62, 0.71], $p < .001$; 7-year-olds: $probability\ of\ success = 0.85$, 95% CI [0.82, 0.89], $p < .001$; binomial exact test). Figure 2 presents the average performance by ratio of proportions and trial type for both age groups.

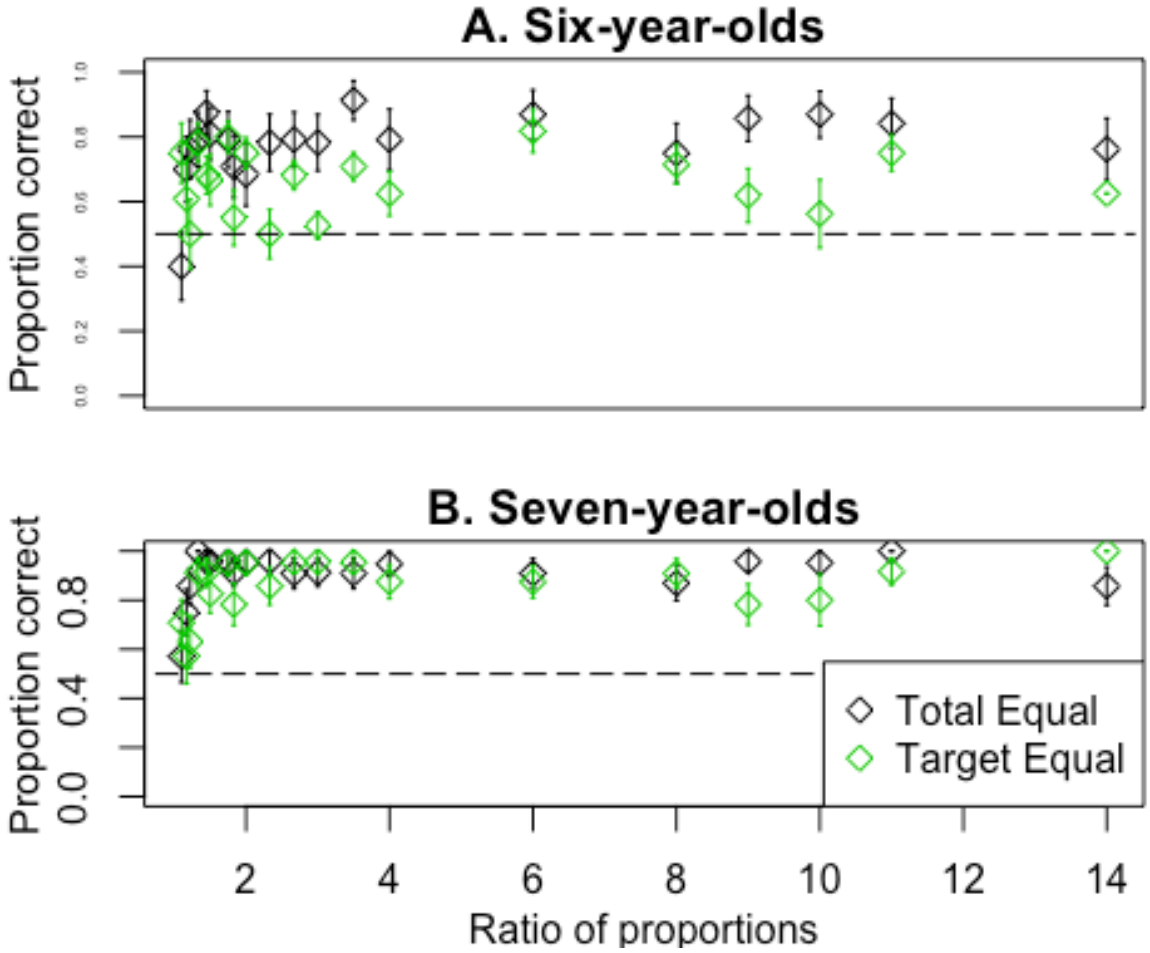

Figure 3.2 Average performance by ratio of proportions and trial type for both 6-year-olds (A) and 7-year-olds (B). Error bars indicate bootstrapped 95% confidence intervals.

Statistical modeling. Generalized Linear Models with Mixed effects (GLMMs) predicted the participant's binary response variable from age, trial type, and ratio of proportions while controlling for the random effects of participant identification number. Preliminary analyses revealed no effects of gender, color of target marble, and order of presentation. For both nested and non-nested models, we use Akaike Information Criterion (AIC) as a method of model selection. AICs are presented alongside the results of chi-square tests of model fit for nested models.

Comparisons of GLMMs revealed that the model with the best fit to the data predicted the participant's response based on trial type, participant age group and the ratio of proportions with no interactions ($AIC_{TT+AG+RP}$ =1,459.66). This model outperformed the null model (AIC_{Null} =1,491.36; χ^2= 37.69; *df* = 3; $p < .001$), the models for trial type (AIC_{TT} =1,472.13; χ^2= 16.47; *df* = 2; $p < .001$) and ratio of proportions (AIC_{RP} =1,486.34; χ^2= 30.67; *df* = 2; $p < .001$), as well as the models based on trial type and age group (AIC_{TT+AG} =1,463.98; χ^2= 6.31; *df* = 1; $p = .01$) and the interaction of trial type and age group (AIC_{TT*AG} =1,465.86). Furthermore, the models which accounted for the interaction between age and ratio of proportions ($AIC_{TT+AG*RP}$ =1,461.12; χ^2= 0.55; *df* = 1; $p = .46$), trial type and ratio of proportions ($AIC_{TT*AG+RP}$ =1,461.37; χ^2= 0.29; *df* = 1; $p = .59$) and the three-way interaction between trial type, age, and ratio of proportions ($AIC_{TT*AG*RP}$ =1,466.07; χ^2= 1.59; *df* = 4; $p = .81$) were not significantly different from the model without interactions. Importantly, these models have a greater number of parameters yet they yield relatively inconsequential improvements in model fits. In this case, the simpler model is preferred because it explains the same amount of variance with fewer parameters.

The preferred model predicts the participant's binary response based on trial type, age group and the ratio of proportions of the presented image ($\beta_{Intercept}$ = 0.75; SE = 0.26; 95% CI [0.24, 1.26]). Inspection of the exponentiated model coefficients revealed that *total equal* trials led to an 85% increase in the odds of obtaining a correct answer (β_{TT} = 0.62; SE = 0.14; 95% CI [0.35, 0.89]). The main effect of age indicated that 7-year-olds performed better than 6-year-olds with the odds of a correct response increasing by a factor of 3.25 for 7-year-olds compared to 6-year-olds (β_{AG} = 1.18; SE = 0.35; 95% CI [0.49, 1.87]). Lastly, we report a main effect of the ratio of proportions with a unit increase in ratio of proportions leading to a 5% increase in the odds of a correct response (β_{RP} = 0.05; SE = 0.02; 95% CI [0.01, 0.09]). Analyses of reaction times yielded similar results details of which can be viewed in the Online Supporting Information.

3.2.3 Discussion

In Experiment 1, our results showed that 6- and 7-year-old children's non-symbolic probability judgments were predicted by the ratio of proportions (i.e., the distance effect). As the ratio of proportions of the two distributions becomes larger, performance improved. These findings are consistent with similar studies with adults (O'Grady, Griffiths, & Xu, 2016; O'Grady et al., submitted) and older children (Fazio et al., 2014). Falk et al. (2012) report that children's probabilistic judgments gradually improve with age and that by the age of 8 children are capable of using the correct proportional strategy. Results from Experiment 1 support these findings. While 7-year-old children performed better than 6-year-old children, both age groups performed worse on *target equal* trials compared to *total equal* trials, indicating that on some trials children may have focused on either the number of target objects or the number of non-target objects without relating the two quantities.

Although the results from Experiment 1 provide novel insight into how young children approximate and reason about binary probabilities, three important design features limit the strength of our findings. First, the images in Experiment 1 consisted of marbles neatly arranged into orderly rows and columns which may have helped some children more accurately approximate the number of marbles in each group. Second, the use of *target equal* trials makes it

difficult to assess whether participants were focusing on the number of target objects or the number of non-target objects. Lastly, all of the marbles in Experiment 1 were the same size yet much of the research using dot approximation tasks has indicated that number approximation is influenced by non-numerical stimulus features such as size and sparsity (Allik, Tuulmets, & Vos, 1991; DeWind, Adams, Platt, & Brannon, 2015; Starr, DeWind, & Brannon, 2017). In order to address these concerns, we designed new stimuli consisting of (1) smaller numbers of marbles randomly positioned on the screen, (2) trials in which the group with a larger proportion of target color marbles contained fewer marbles of the target color than the group with the smaller proportion of target color marbles (3) trials in which the target marbles in the 'losing' distributions were larger than the target marbles in the 'winning' distribution. Since we expected each of these changes to increase the difficulty of the task and performance of the 6-year-olds in Experiment 1 was already relatively low, we decided to test older children for Experiment 2.

3.3 Experiment 2

3.3.1 Methods

Participants. One hundred and forty-two children between the ages of 7 and 12 were recruited from local schools and children's museums from the San Francisco Bay area. Twelve of these children were excluded from our analyses: eight children were excluded due to experimenter error, three children did not pass the practice trials, and one child's parent interfered in the study by coaching their child to choose the group with a larger proportion of target marbles. As with Experiment 1 our target sample size was determined based on previous research (Fazio et al., 2014). However, since our target age range was much larger (7- to 12-year-olds), data collection continued according to a stopping rule requiring a minimum of 40 participants in each of three age groups (7- to 8-year-olds, 9- to 10-year-olds, and 11- to 12-year-olds). The final sample consisted of forty 7- and 8-year-olds (N = 40; Mean age = 7.96; SD = 0.53; 24 female), fifty 9- and 10-year-olds (N = 50; Mean age = 10.04; SD = 0.50; 20 female), and forty 11- and 12-year-olds (40; Mean age = 11.75; SD = 0.64; 18 female). Data collection was conducted in the same schools and communities reported in Experiment 1.

Material. As mentioned above, the orderly arrangement of marbles in Experiment 1 may have helped children approximate the number of marbles in each group. In order to prevent this, the location of each marble was randomly generated for each image using Blender 2.72. Due to the ceiling levels of performance for high ratios of proportions in Experiment 1 we decided to include more trials with lower ratios of proportions, ranging from 1.1 (55% vs. 50%) to 9.5 (95% vs. 10%). We also decided to include ratios of two proportions that were both below chance (i.e. 40% to 15%). Table 2 presents the proportions of marbles in each group alongside the ratios of proportions used in Experiment 2. For each ratio of proportions, three trial types were created: *total equal* trials in which each distribution had the same total number of marbles; *area-anticorrelated* trials in which the sizes of the marbles were manipulated such that the total area covered by the target marbles in the 'losing' group was larger than the total area covered by the target marbles in the 'winning' group. Importantly, *area-anticorrelated* trials included groups of marbles with an equal total amount of marbles similar to *total equal* trials. Finally, *number vs. proportion* trials in which the distribution with the lower proportion of target marbles contained a

larger number of target marbles. A total of 264 Images were rendered and then divided equally into four conditions based on target color and the order of presentation of the images. Each participant viewed 66 images presented in one of four conditions (Red, order 1; Red, order 2; White, order 1; and White, order 2). Importantly, the order of the images was pseudorandomized such that there were no more than 3 consecutive trials in which the 'correct' choice was on the same side of the screen.

Table 3.2 Proportions presented in each trial of Experiment 2.

Proportion group 1	Proportion group 2	Ratio of Proportions
0.55	0.50	1.10
0.50	0.45	1.11
0.45	0.40	1.12
0.55	0.45	1.22
0.80	0.60	1.33
0.80	0.55	1.45
0.75	0.50	1.50
0.60	0.40	1.50
0.95	0.55	1.73
0.95	0.50	1.90
0.50	0.25	2.00
0.40	0.15	2.67
0.60	0.20	3.00
0.50	0.15	3.33
0.80	0.20	4.00
0.40	0.10	4.00
0.75	0.15	5.00
0.95	0.15	6.33
0.80	0.10	8.00
0.85	0.10	8.50
0.90	0.10	9.00
0.95	0.10	9.50

Note. Ratios of Proportions are rounded to 2 decimal points.

Procedure. After parental guardians provided written consent for their children to participate in the study, children were asked to provide verbal assent and were then seated in front of a MacBook Pro laptop (OSX; Screen resolution 1280 x 800). Participants were told that they were going to play a game in which they would collect red or white marbles depending on the condition to which they were assigned. Children were shown two boxes and told that they would see two groups of marbles on two trays on the screen. The group of marbles on the left side of

the screen were poured into the box on the left and the group on the right side of the screen were poured into the box on the right side of the screen. The boxes would then be shaken up so that they could not infer the positions of the marbles based on their locations on the viewing trays. They were then asked to select the box that they thought was best for collecting their target color marble. After the instructions phase participants played 4 practice trials in order to ensure that they understood the game. Once the practice trials were complete, participants were presented with 66 semi-randomized test trials in which they were able to view the images for 1500 ms before making their selection. As with Experiment 1 short intermission videos were played after the 15th, 30th, and 45th trials, children were not given any feedback about their decisions, and the experimenter never mentioned the words 'probability' or 'proportion'. Figure 3 provides a visual schematic of the procedure.

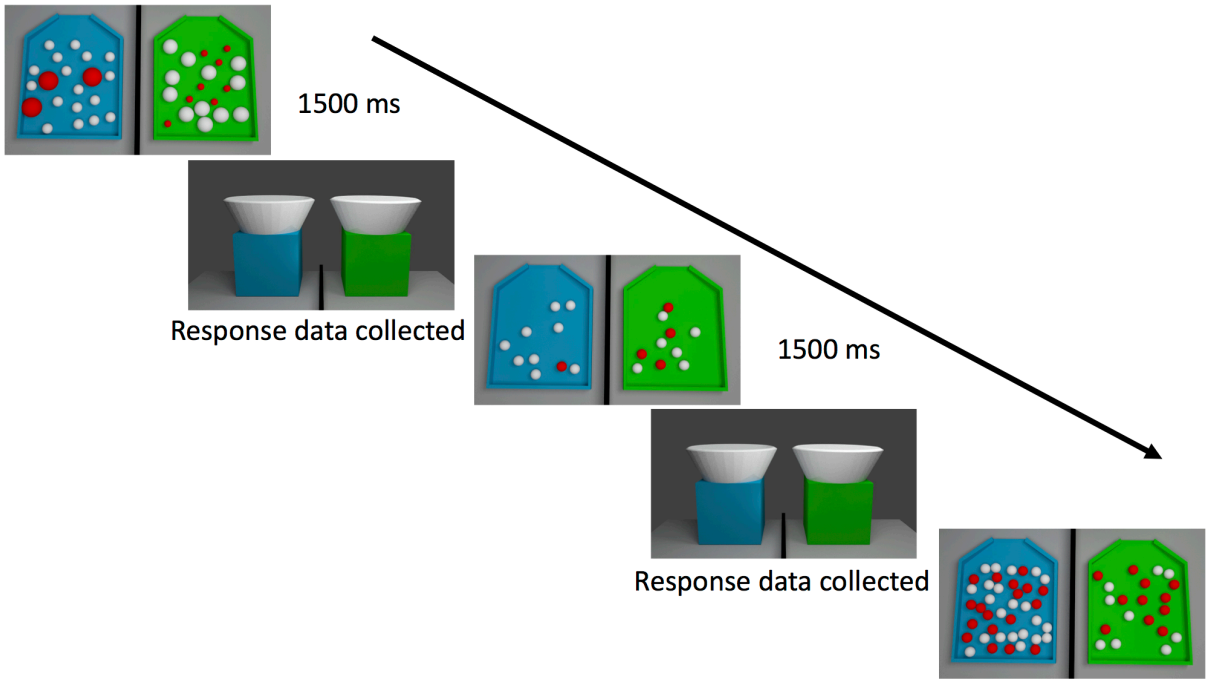

Figure 3.3 Diagram of the experimental procedure used in Experiment 2. The sample image at the top presents an *area anti-correlated* trial, the sample image in the middle presents a *total equal* trial and the sample image at the bottom presents a *number vs. proportion* trial.

3.3.2 Results

Reaction time. The same procedure employed in Experiment 1 for cleaning outlying reaction time resulted in the exclusion of 1383 out of 8580 trials (16.12%). As with Experiment 1, exclusion of trials with outlying reaction time does not change the reported results. A report of the results of the same analyses for the complete dataset for Experiment 2 are included in the Online Supporting Information. Of the included trials, reaction times were significantly lower on *total equal* trials (M = 815.29 ms, SD = 466.83) and *number vs. proportion* trials (M = 819.06 ms, SD = 441.59) compared to *area-anticorrelated* trials (M = 888.15 ms, SD = 480.37; total equal: $\Delta M = 72.86$, 95% CI $[-99.77, -45.95]$, $t(4{,}758.99) = -5.31$, $p < .001$; number vs. proportion: $\Delta M = 69.09$, 95% CI $[-95.08, -43.11]$, $t(4{,}805.06) = -5.21$, $p < .001$). The

difference between *total equal* trials and *number vs. proportion* trials did not reach significance ($\Delta M = 3.77$, 95% CI $[-29.55, 22.00]$, $t(4{,}741.64) = -0.29$, $p = .774$).

General accuracy. Results of the binomial exact tests comparing performance against chance revealed that children in all three age groups performed significantly above chance on both *total equal* trials (8-year-olds: $probability\ of\ success = 0.78$, 95% CI $[0.75, 0.81]$, $p < .001$; 10-year-olds: $probability\ of\ success = 0.87$, 95% CI $[0.85, 0.89]$, $p < .001$; 12-year-olds: $probability\ of\ success = 0.91$, 95% CI $[0.88, 0.93]$, $p < .001$) and *area-anticorrelated* trials (8-year-olds: $probability\ of\ success = 0.60$, 95% CI $[0.57, 0.64]$, $p < .001$; 10-year-olds: $probability\ ofs\ uccess = 0.69$, 95% CI $[0.66, 0.72]$, $p < .001$; 12-year-olds: $probability\ of\ success = 0.61$, 95% CI $[0.57, 0.64]$, $p < .001$). Finally, 12-year-old and 10 -year-old children performed significantly better than chance on *number vs. proportion* trials (10-year-olds: $probability\ of\ success = 0.55$, 95% CI $[0.52, 0.58]$, $p = .003$; 12-year-olds: $probability\ of\ success = 0.61$, 95% CI $[0.57, 0.64]$, $p < .001$) while 8-year-olds' performance was not significantly different from chance on *number vs. proportion trials* (8-year-olds: $probability\ of\ success = 0.46$, 95% CI $[0.43, 0.50]$, $p = .056$). Figure 4 presents the proportion of correct responses by ratio of proportions, trial type, and age group.

Statistical modeling. As with Experiment 1, we compared Generalized Linear Models with Mixed effects (GLMM) and used Akaike Information Criterion (AIC) as our method for model selection for non-nested models and chi-square tests for nested models. Results of the model comparisons revealed that the model predicting performance from the 3-way interaction between trial type, ratio of proportions, and age group ($AIC_{FullModel}$ =7,412.06) outperformed all other models including the null model (AIC_{Null} =8,503.15; χ^2= 1,125.10; *df* = 17; *p* < .001), the models for trial type (AIC_{TT} =7,887.38; χ^2= 505.32; *df* = 15; *p* < .001), ratio of proportions (AIC_{RP} =8,147.82; χ^2= 767.76; *df* = 16; *p* < .001) as well as more complex models based on trial type and age group (AIC_{TT+AG} =7,864.39; χ^2= 478.33; *df* = 13; *p* < .001), trial type and ratio of proportions (AIC_{TT+RP} =7,488.82; χ^2= 104.77; *df* = 14; *p* < .001), the interaction of trial type and ratio of proportions ($AIC_{TT*RP+AG}$ =7,444.83; χ^2= 80.47; *df* = 12; *p* < .001) and the interaction of trial type and age group ($AIC_{TT*AG+RP}$ =7,444.83; χ^2= 68.53; *df* = 8; *p* < .001).

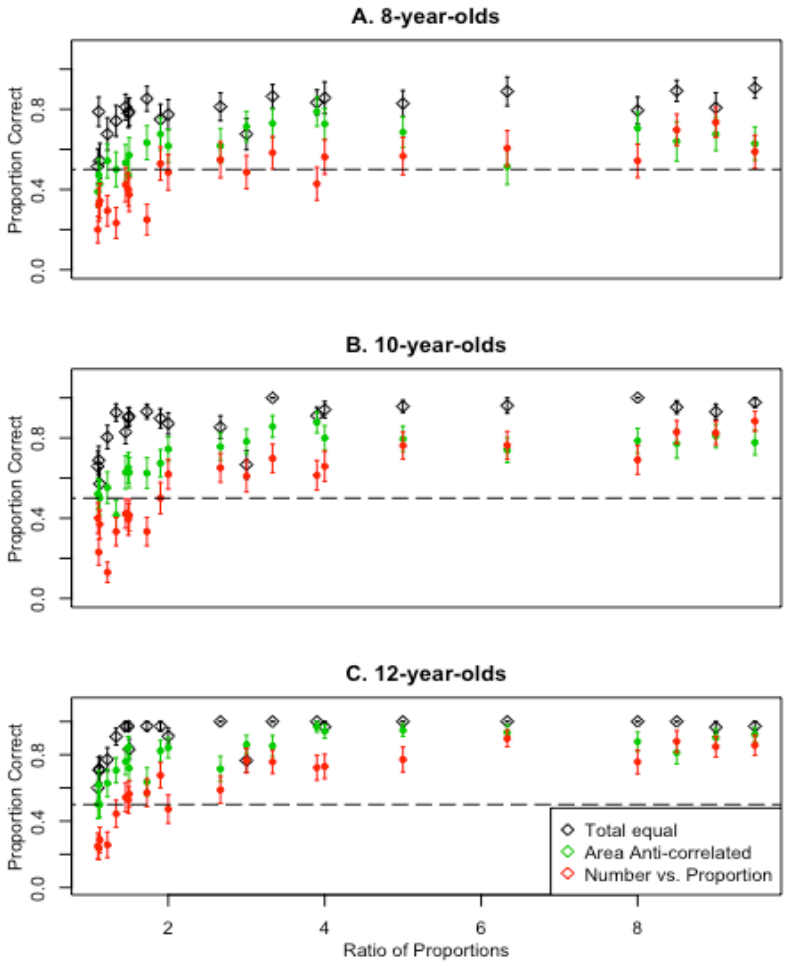

Figure 3.4 Average performance by the log of ratio of proportions and trial type. A) 8-year-olds. B) 10-year-olds. C) 12-year-olds. Error bars indicate bootstrapped 95% Confidence Intervals.

Coefficients in logistic regression indicate the change in log-odds of a correct response based on changes in experimental and subject variables. It is easiest to express these changes by exponentiating the coefficients to reveal the change in odds and to relate these changes to the baseline group: 8-year-olds' responses to *total equal* images ($\beta_{Intercept}$ = 0.88). Exponentiated model coefficients for main effects of trial type revealed *Number vs. proportion* trials led to an 82% reduction in the odds of a correct response (β_{NvP} = -1.73) while *area anticorrelated* trials only lead to a 52% decrease in the odds (β_{AA} = -0.75). Main effects of age indicated that performance improved with each age group (β_{Age10} = 0.25; β_{Age12} = 0.21). Since the ratio of proportions is a continuous variable, the associated coefficient revealed that a single incremental change in ratio of proportions led to a 15% increase in the odds of a correct response (β_{RP} = 0.14). The only interaction term to reach significance indicated that the effect of ratio of proportions increased for 12-year-olds compared to the younger children ($\beta_{Age12XRP}$ = 0.14). The full set of fixed effect model coefficients are presented in Table 4 in the Online Supporting

Information. Analyses of reaction time data from Experiment 2 revealed a similar pattern of results, the details of which are available in the Online Supporting Information.

3.3.3 Discussion

Experiment 2 replicated the main results of Experiment 1 and extended these findings by including three trial types (i.e., total equal, number vs. proportion, and area-anticorrelated) and three age groups. As in Experiment 1, we found that the performance of children of all three age groups was strongly influenced by the ratio of proportions, converging with results from adults using a very similar methodology (O'Grady, Griffiths, & Xu, submitted) and those of Fazio et al. (2014) with 12-year-old children. By including area-anticorrelated trials, our results revealed that children relied on both numerical and non-numerical stimulus features; that is, children made more errors when the total area of the target marbles was larger in one distribution even when the proportion of target marbles was smaller in that distribution. The model coefficient for *area anticorrelated* trials suggests that children of all ages were influenced by this manipulation, indicating that non-numerical stimulus features influence probability judgments. By including number vs. proportion trials, our results revealed that children up to 10 often used a formally incorrect strategy in estimating proportions and probability; that is, they used the number of target marbles in a distribution as a proxy for estimating the proportion of target marbles. The only age group that performed above chance level on these trials was the oldest (12-year-olds), and their accuracy was far from perfect (only 60%).

3.4 General Discussion

In two experiments, we provide evidence that 6- to 12-year-old children can make rapid and accurate approximations of probability based on proportions. Our findings are consistent with the results of other ratio comparison tasks with adults (O'Grady et al., submitted), 12-year-old children (Fazio et al., 2014) and non-human primates (Drucker, Rossa, & Brannon, 2016). More importantly, our results shed new light on the development of proportional and probabilistic reasoning. In Experiment 1 we report data demonstrating that 6- and 7-year-olds' non-symbolic probability judgments are characterized by ratio dependence and that the acuity of these representations improves with age. Experiment 2 replicated these findings and also revealed that non-symbolic probability judgments are influenced by the same numerical and non-numerical stimulus features which influence perceived numerosity such as the size of dots in a dot discrimination task. Data from both experiments also suggests that children produced similar errors in our non-symbolic probability estimation task compared previous research using the 2AFC random draw task (Falk et al. 2012) as well as studies on children's proportional reasoning (Boyer, Levine & Huttenlocher, 2008; Hurst & Cordes, 2018). More specifically, children's performance was influenced by the number of target marbles as evidenced by their decreased performance on *number vs proportion* trials relative to total equal trials.

These findings make three important contributions to the literature on probabilistic reasoning, proportional reasoning, and quantitative development in general. First, we provide evidence that children's probability judgments are characterized by ratio dependence and even young children can make accurate judgments about the likelihood of future events based on proportions. Second, our current experiments represent the first attempt to systematically investigate the mental representation and psychophysical properties of non-symbolic probability

in the developing human mind. We provide evidence that school-age children's probability estimation is influenced by the size of the objects being approximated. Third, previous research has not charted the developmental trajectory of the mental representation of non-symbolic probability. We provide the first evidence that between 6 and 12, children's ability to estimate probability improves with age, and they gradually adopt the correct proportion strategy although they continue to make errors by sometimes employing heuristic decision rules.

The results of the current experiments provide new insights on the role of proportional reasoning in children's probability judgments; they also raise important questions for future research. In the current studies, 9- to 12-year-old children performed above chance in the number vs. proportion trials, whereas the 8-year-olds did not. In contrast, there is some evidence suggesting that both infants and non-human primates can use ratio of proportions in estimating probability (Denison & Xu, 2014, Rakoczy et al. 2014). In one study, infant participants observed an experimenter randomly draw a single lollipop from each of two groups of preferred and non-preferred color lollipops (Denison & Xu, 2014). Infants were more likely to approach the lollipop drawn from the distribution with a larger proportion of preferred lollipops even when the total number of lollipops in both groups varied such that the group with the lower proportion actually contained more of the infant's preferred lollipops. It may be the case that ANS acuity improves with age, and the current studies used ratios of proportions that were more difficult than that of Denison and Xu (2014). However, in Experiment 2 of Denison and Xu (2014), infants performed above chance when presented a ratio of proportions of 4 (80% target objects in one distribution vs. 20% target objects in the other distribution). The current studies included the same ratio of proportions yet it is not until about 9 that children succeeded on the number vs. proportion trials. One possible explanation for this is that older children's poor performance on these trials may be due to the 'whole number bias' reported in the education literature on rational number learning (Ni & Zhou, 2005). In the fraction learning literature, the 'whole number bias' is observed most often when children choose the larger of two fractions based on the magnitude of the components of the fractions (i.e. by choosing the fraction with the larger numerator or denominator) rather than selecting the larger fraction based on the relation between numerator and denominator. The literature on probability reasoning has investigated this same response bias in the context of probability predictions beginning with the seminal work of Piaget & Inhelder (1975) and recent work (Falk et al. 2012) has indicated that this type of response bias constitutes a strategy that younger children use in 2AFC probability judgment tasks.

The integrated theory of numerical development (Siegler, 2016; Siegler, Thompson, & Schneider, 2011) posits that children come to understand rational numbers through analogy to whole numbers and evidence from studies on proportional reasoning suggests that children overextend their knowledge of whole numbers when reasoning about proportions presented as discretized units rather than continuous quantities (Boyer, Levine, & Huttenlocher, 2008; Mix, Huttenlocher, & Levine, 2002). It is possible that an overreliance on whole number knowledge led to younger children's incorrect choices on *number vs. proportion* trials. To explore this possibility, we are currently developing a modified version of our probability discrimination task for use with much younger, preschool-age children and toddlers. The prediction is that much younger children may succeed in using proportions to estimate probability (consistent with the findings with infants) whereas older, school-age children may adopt the whole number strategy. Indeed, preliminary evidence (O'Grady & Xu, 2018) has shown that school age children demonstrate a whole number bias when making probability judgment tasks involving both exact

and approximate quantities but this bias can be overridden if the child is provided with enough feedback.

Our experiments also raise new questions about the role of magnitude processing in proportional reasoning and probabilistic estimation. The current body of literature suggests two possibilities. One is that the Approximate Number System serves as a building block for computing probabilities. According to this account, children first approximate the number of marbles of each type within each group and then use these approximate representations to compute the probabilities. Specifically, probabilities are computed as follows: (number of target objects) / (number of target objects + number of non-target objects). A second possibility suggests that children bypass discrete number approximations altogether and simply approximate ratios using a Ratio Processing System (RPS). Recent research has provided a wealth of evidence to suggest that ratio processing is fundamental to human numerical cognition (Matthews & Chesney, 2015; Matthews & Lewis, 2017), and thus constitutes a basic building block for learning symbolic fractions (Matthews, Lewis, & Hubbard, 2015). While we agree that ratio processing is foundational for mathematics learning, it is unclear whether the RPS and the ANS are two separable systems. Indeed researchers studying early numerical development have recently argued that there exists a general magnitude processing system in the brain, that includes estimations of number (integers, proportions, and probability), duration, and spatial extent (Mix, Levine, & Newcombe, 2016; Lourenco, 2016). Thus, we tend to favor the former claim that children draw on ANS representations for three reasons. First, it is unclear at the moment whether the RPS exists independently of the ANS. Second, in order to calculate the probability of a discrete event, decision-makers must represent discrete outcomes. It is possible that the RPS may be able to compute proportions of discrete elements and this is exactly the type of argument that would support the notion that RPS and ANS are two elements of a more generalized magnitude processing system. Indeed, Jacob, Valentin, & Nieder (2012) have suggested that the ANS may provide one source of input for the RPS. Third, children's performance on number vs proportion trials in our experiment suggests that number approximations may play an important role in their probability judgments. This claim is clearly speculative based on the current series of studies, but it provides an important avenue for future research and the domain of probabilistic reasoning offers an interesting way to study the relationship between the ANS and RPS.

Lastly, the current studies are also limited in the types of probabilistic reasoning they address. Based on the findings in the proportional reasoning literature (Boyer et al., 2008), a natural extension of the current work is to investigate whether children rely on the same heuristic decision rules we find for discrete probability (i.e. marbles drawn from a container) when making judgments about continuous probability (i.e. spinner tasks). Furthermore, the current set of experiments focus exclusively on simultaneously presented visual information, thus the role of number and ratio approximation in judgments about sequentially presented probability problems, similar to the methods reported in Boyer (2007), cannot be addressed by the current findings. Future work will investigate whether children and adults rely on ratio processing and integer approximation when tracking and computing the probability of sequentially presented stimuli.

These findings indicate that children can make rapid estimations about the probability of discrete outcomes. Furthermore, we have shown that these representations share some common features with perceptual systems for processing numerical magnitude. By linking the developmental literatures on the approximate number system and probabilistic reasoning we have demonstrated children's intuitive ability to estimate probability is surprisingly accurate.

While our results are perhaps most relevant to researchers and educators studying the development of numerical cognition and quantitative development, they may also inform research from a variety of subfields in developmental psychology, such as the development of decision-making strategies and scientific reasoning.

All methods, analyses, and de-identified data are available on the Open Science Framework (https://osf.io/48sgv/).

3.5 Appendix

3.5.1 Additional information for Experiment 1

Reaction time analyses for Experiment 1. Reaction time data in Experiments 1 and 2 were analyzed using comparisons of Linear regression Models with Mixed effects (LMMs). This analytical method is mathematically equivalent to a repeated measures anova and entering participant identification number as a random effect allows us to account for repeated measures. As with accuracy, model comparisons did not reveal any effect of gender or the order of trial presentation. However, a model predicting performance based on the color of the target marbles (AIC_{Color} =27,286.84) outperformed the null model (AIC_{Null} =27,288.98; χ^2= 4.14; df = 1; p < .04). Analyses of the model coefficients ($\beta_{Intercept}$ = 774.09; SE = 124.75; 95% CI [529.59, 1,018.59]) revealed that on average, participants were slower when the color of the target marble was white compared to when target color was red (β_{Color} = 352.64; SE = 124.75; 95% CI [20.32, 684.95]). It is possible the contrast between the red marbles and the gray background color was a contributing factor to this effect. However, since the current experiment was not designed to assess the influence of color on number approximation, it is difficult to make any strong conclusions about the effect of target marble color.

Comparisons of LMMs for reaction time data revealed that the model predicting reaction time from trial type, age group, ratio of proportions and the interaction between trial type and age group was best fit to the data ($AIC_{TTXAG+RP}$ =27,173.51). This model outperformed the null model (AIC_{Null} =27,288.98; χ^2= 63.51; df = 4; p < .001) as well as the models for trial type (AIC_{TT} =27,257.27; χ^2= 29.80; df = 3; p < .001), age group (AIC_{AG} =27,279.89; χ^2= 52.42; df = 3; p < .001), and ratio of proportions (AIC_{RP} =27,279.11; χ^2= 51.64; df = 3; p < .001). Furthermore, this model outperformed the model predicting reaction time from trial type, age group, ratio of proportions and the interaction between age group and ratio of proportions ($AIC_{TT+AGXRP}$ =27,173.51) as well as the model based on all three variables without any interactions ($AIC_{TT+AG+RP}$ =27,236.15; χ^2= 4.68; df = 1; p < .03). Importantly, the full model which can account for interactions between the three factors was not a better fit to the data even though it had more parameters ($AIC_{FullModel}$ =27,237.89; χ^2= 4.41; df = 0; p < .00). Model coefficients along with standard error are presented in table 1 below.

Table 3.A1 Coefficients for fixed effects of best fit model for reaction time data from Experiment 3.1

	Est. Coefficient	SE	L 95% CI	U 95% CI
Intercept	1,306.52	116.81	1,077.57	1,535.47
Target equal	-247.99	42.90	-332.07	-163.90
Age 7	-624.12	163.36	-944.31	-303.93
RP	15.72	4.21	7.47	23.97
Target equal X Age 7	130.69	60.44	12.23	249.15

Note. All values are rounded to the second decimal.

Analyses for the complete data set in Experiment 1. In this section we report the results of the regression analyses without excluding trials based on outlying reaction times.

General accuracy. Performance on foil trials was above chance for both 6-year-olds ($M = 0.79$, 95% CI [0.69, 0.88], $t(23) = 6.25$, $p < .001$) and 7-year-olds ($M = 0.97$, 95% CI [0.91, 1.03], $t(23) = 15.94$, $p < .001$) suggesting that children did not learn to merely select the smallest group when presented with groups of different sizes. Foil trials were excluded from the remainder of analyses. Children in both age groups performed significantly above chance on both *total equal* (6-year-olds: $M = 0.76$, 95% CI [0.68, 0.85], $t(23) = 6.42$, $p < .001$; 7-year-olds: $M = 0.91$, 95% CI [0.86, 0.96], $t(23) = 17.87$, $p < .001$) and *target equal* trial types (6-year-olds: $M = 0.66$, 95% CI [0.58, 0.74], $t(23) = 4.09$, $p < .001$; 7-year-olds: $M = 0.84$, 95% CI [0.78, 0.90], $t(23) = 11.31$, $p < .001$).

Statistical modeling. Comparisons of GLMMs revealed that the model with the best fit to the data predicted the participant's response based on trial type, participant age group and the ratio of proportions with no interactions ($AIC_{TT+AG+RP}$ =1,725.08). This model outperformed the null model (AIC_{Null} =1,768.71; χ^2= 49.63; *df* = 3; $p < .001$), the models for trial type (AIC_{TT} =1,743.38; χ^2= 22.30; *df* = 2; $p < .001$) and ratio of proportions (AIC_{RP} =1,762.39; χ^2= 41.31; *df* = 2; $p < .001$), as well as the models based on trial type and age group (AIC_{TT+AG} =1,731.55; χ^2= 8.47; *df* = 1; $p = .00$) and the interaction of trial type and age group (AIC_{TT*AG} =1,733.36). Furthermore, the models which accounted for the interaction between age and ratio of proportions ($AIC_{TT+AG*RP}$ =1,726.61; χ^2= 0.47; *df* = 1; $p = .49$), trial type and ratio of proportions ($AIC_{TT*AG+RP}$ =1,726.53; χ^2= 0.55; *df* = 1; $p = .46$) and the three-way interaction between trial type, age, and ratio of proportions ($AIC_{TT*AG*RP}$ =1,731.73; χ^2= 1.35; *df* = 4; $p = .85$) were not significantly different from the model without interactions. Importantly, these models have a greater number of parameters yet they yield relatively inconsequential improvements in model fits. In this case, the simpler model is preferred because it explains the same amount of variance with fewer parameters. The preferred model predicts the participant's binary response based on trial type, age group and the ratio of proportions of the presented image ($\beta_{Intercept}$ = 0.55; SE = 0.22; 95% CI [0.12, 0.98]). Inspection of the exponentiated model coefficients revealed that *total equal* trials led to an 85% increase in the odds of obtaining a correct answer (β_{TT} = 0.65; SE = 0.12; 95% CI [0.41, 0.90]). The main effect of age indicated

that 7-year-olds performed better than 6-year-olds with the odds of a correct response increasing by a factor of 3.25 for 7-year-olds compared to 6-year-olds (β_{AG} = 1.16; SE = 0.29; 95% CI [0.58, 1.73]). Lastly, we report a main effect of the ratio of proportions with a unit increase in ratio of proportions leading to a 5% increase in the odds of a correct response (β_{RP} = 0.05; SE = 0.02; 95% CI [0.02, 0.08]).

3.5.2 Additional information for Experiment 2

Reaction time analyses for Experiment 2. Comparisons of LMMs for reaction time data in Experiment 2 revealed that the full model predicting reaction time from trial type, ratio of proportions, and age as well as all interactions had the best fit to the data. Table 2 presents the results of the model comparisons and table 3 presents the model coefficients

Table 3.A2 Model comparisons for the data from Experiment 3.2

	Model	AIC	df	Chi-squared	P-value
AIC	Full Model	105797.617886164	17		
AIC	Null Model	105916.74340409	17	153.125517926004	< .001
AIC	Trial Type (TT)	105858.218836892	15	90.6009507276904	< .001
AIC	Age Group (AG)	105897.720538889	15	130.10265272473	< .001
AIC	Ratio of Proportions (RP)	105888.678789496	16	123.060903331818	< .001
AIC	TT+AG+RP	105811.415410916	8	33.5852298987011	< .001
AIC	TTRPintAG	105815.294960352	10	37.6770741874352	< .001
AIC	TTAGintRP	105815.203116063	8	33.5852298987011	< .001

Note. All values are rounded to the second decimal.

Table 3.A3 Coefficients for fixed effects of the full model for data from Experiment 3.2

	Est. Coefficient	SE	L 95% CI	U 95% CI
Intercept	898.92	46.50	807.78	990.07
Number vs Proportion (NvP)	14.96	31.19	-46.17	76.09
Area Anti-correlated (AA)	58.95	31.16	-2.13	120.04
Age 10	-162.79	62.05	-284.40	-41.18
Age 12	-206.00	65.34	-334.07	-77.94
Ratio of Proportions (RP)	19.97	4.97	10.23	29.71
NP X Age 10	-19.92	41.72	-101.69	61.86
AA X Age 10	6.41	41.76	-75.45	88.27
NvP X Age 12	-18.07	43.70	-103.73	67.60
AA X Age 12	25.06	43.90	-60.98	111.10
NP X RP	-6.50	6.92	-20.06	7.07
AA X RP	7.85	6.96	-5.79	21.50
Age 10 X RP	-12.85	6.58	-25.75	0.05
Age 12 X RP	-22.09	6.98	-35.77	-8.42
NvP X Age 10 X RP	5.25	9.24	-12.86	23.35
AA X Age 10 X RP	-10.86	9.27	-29.02	7.31
NvP X Age 12 X RP	14.41	9.75	-4.70	33.51
AA X Age 12 X RP	-8.06	9.78	-27.22	11.10

Note. All values are rounded to the second decimal.

Table 3.A4 Coefficients for fixed effects of the full model for accuracy data of Experiment 3.2

	Est. Coefficient	SE	L 95% CI	U 95% CI	Wald Z	p value
Intercept	0.88	0.19	0.51	1.25	4.70	< .001
Number vs. Proportion (NvP)	-1.73	0.20	-2.12	-1.33	-8.63	< .001
Area Anti-correlated (AA)	-0.75	0.20	-1.14	-0.37	-3.81	< .001
Age 10	0.25	0.27	-0.28	0.79	0.92	.36
Age 12	0.21	0.35	-0.48	0.90	0.58	.56
Ratio of Proportions (RP)	0.14	0.04	0.06	0.22	3.62	< .001
NP X Age 10	-0.30	0.29	-0.87	0.28	-1.01	.31
AA X Age 10	-0.02	0.29	-0.58	0.55	-0.06	.96
NvP X Age 12	0.02	0.37	-0.70	0.74	0.06	.95
AA X Age 12	0.33	0.37	-0.40	1.06	0.89	.37
NP X RP	0.04	0.05	-0.05	0.14	0.89	.37
AA X RP	-0.05	0.05	-0.15	0.04	-1.07	.28
Age 10 X RP	0.18	0.07	0.04	0.32	2.57	.01
Age 12 X RP	0.45	0.14	0.18	0.72	3.26	< .001
NvP X Age 10 X RP	-0.04	0.08	-0.20	0.13	-0.43	.67
AA X Age 10 X RP	-0.11	0.08	-0.27	0.05	-1.33	.18
NvP X Age 12 X RP	-0.29	0.15	-0.58	0.00	-1.98	.05
AA X Age 12 X RP	-0.28	0.15	-0.57	0.01	-1.88	.06

Note. All values are rounded to the second decimal.

Analyses for the complete data set in Experiment 2

In this section we report the results of the regression analyses without excluding trials based on outlying reaction times.

General accuracy

Children in all three age groups performed significantly above chance on *total equal* trials (8-year-olds: $M = 0.78$, 95% CI [0.72, 0.83], $t(39) = 10.06$, $p < .001$; 10-year-olds: $M = 0.86$, 95% CI [0.82, 0.90], $t(49) = 19.70$, $p < .001$; 12-year-olds: $M = 0.90$, 95% CI [0.88, 0.93], $t(39) = 31.69$, $p < .001$). Interestingly, 10 and 12 year olds performed significantly above chance on *area-anticorrelated* trials (10-year-olds: $M = 0.68$, 95% CI [0.60, 0.75], $t(49) = 4.61$, $p < .001$; 12-year-olds:$M = 0.78$, 95% CI [0.73, 0.84], $t(39) = 10.07$, $p < .001$) while 8-year-olds' performance did not differ from chance for area-anticorrelated trials (8-year-olds: $M = 0.58$, 95% CI [0.49, 0.66], $t(39) = 1.84$, $p = .074$). Finally, 12-year-old children performed significantly better than chance on *number vs. proportion* trials (12-year-olds: $M = 0.60$, 95% CI [0.51, 0.70], $t(39) = 2.24$, $p = .031$) while the two younger age groups demonstrated performance at chance levels (8-year-olds: $M = 0.48$, 95% CI [0.40, 0.56], $t(39) = -0.61$, $p = .545$; 10-year-olds: $M = 0.54$, 95% CI [0.47, 0.62], $t(49) = 1.26$, $p = .212$).

Statistical modeling

Results of the model comparisons for the full set of data revealed that the model predicting performance from the 3-way interaction between trial type, ratio of proportions, and age group ($AIC_{FullModel}$ =8,951.83) outperformed all other models including the null model (AIC_{Null} =10,190.73; χ^2= 1,272.90; *df* = 17; $p < .001$), the models for trial type (AIC_{TT} =9,474.77; χ^2= 552.94; *df* = 15; $p < .001$), ratio of proportions (AIC_{RP} =9,786.64; χ^2= 866.81; *df* = 16; $p < .001$) as well as more complex models based on trial type and age group (AIC_{TT+AG} =9,451.71; χ^2= 525.88; *df* = 13; $p < .001$), trial type and ratio of proportions (AIC_{TT+RP} =9,032.36; χ^2= 108.53; *df* = 14; $p < .001$), the interaction of trial type and ratio of proportions ($AIC_{TT*RP+AG}$ =8,988.75; χ^2= 84.00; *df* = 12; $p < .001$) and the interaction of trial type and age group ($AIC_{TT*AG+RP}$ =8,988.75; χ^2= 66.46; *df* = 8; $p < .001$).

Coefficients of the model for the full set of data were of a similar magnitude and direction to the corresponding model based on the data when excluding trials with outlying reaction times ($\beta_{Intercept}$ = 0.89, β_{NvP} = -1.61, β_{AA} = -0.89, β_{Age10} = 0.18; β_{Age12} = 0.29, β_{RP} = 0.14, $\beta_{Age12XRP}$ = 0.14). The full set of model coefficients for the complete data set are reported in Table 5.

Table 3.A5 Coefficients for fixed effects of the full model for the complete dataset from Experiment 3.2

	Est. Coefficient	SE	L 95% CI	U 95% CI	Wald Z	p value
Intercept	0.89	0.17	0.55	1.23	5.13	< .001
Number vs. Proportion (NvP)	-1.61	0.18	-1.96	-1.25	-8.89	< .001
Area Anti-correlated (AA)	-0.89	0.18	-1.24	-0.54	-4.97	< .001
Age 10	0.18	0.25	-0.32	0.67	0.70	.49
Age 12	0.29	0.30	-0.30	0.88	0.96	.33
Ratio of Proportions (RP)	0.14	0.03	0.07	0.21	4.10	< .001
NP X Age 10	-0.32	0.26	-0.84	0.19	-1.23	.22
AA X Age 10	0.11	0.26	-0.40	0.62	0.43	.67
NvP X Age 12	-0.09	0.31	-0.70	0.52	-0.29	.78
AA X Age 12	0.39	0.32	-0.23	1.01	1.23	.22
NP X RP	0.03	0.04	-0.06	0.11	0.67	.50
AA X RP	-0.05	0.04	-0.13	0.04	-1.07	.29
Age 10 X RP	0.18	0.06	0.06	0.30	2.85	< .001
Age 12 X RP	0.34	0.10	0.14	0.54	3.39	< .001
NvP X Age 10 X RP	-0.04	0.07	-0.18	0.11	-0.50	.62
AA X Age 10 X RP	-0.12	0.07	-0.26	0.02	-1.63	.10
NvP X Age 12 X RP	-0.21	0.11	-0.42	0.00	-1.95	.05
AA X Age 12 X RP	-0.20	0.11	-0.42	0.02	-1.78	.07

Note. All values are rounded to the second decimal.

Chapter 4

Strategy-specific Feedback Influences Children's Use of Heuristics In Probability Judgment Tasks.

4.1 Introduction

The uncertain nature of chance and probability underlies nearly every instance of human decision making. Whether we are deciding which career paths to pursue or what to wear to work in the morning, our decisions are informed by considering probabilistic data (i.e. job placement rates or weather forecasts) as well as our understanding of formal probability. For decades researchers have claimed that humans are impoverished reasoners of chance, presenting data that suggest even educated adults make poor choices based on probabilistic data (Kahneman, 2011; Kahneman & Tversky, 1973; Tversky & Kahneman, 1983). However, a wealth of recent evidence suggests that infants (Denison & Xu, 2014), apes (Rakoczy et al., 2014), capuchin monkeys (Tecwyn, Denison, Messer, & Buchsbaum, 2017), and rhesus macaques (De Petrillo & Rosati, 2019) are capable of intuitive probabilistic reasoning, suggesting that the human ability to accurately reason about uncertainty develops early in ontogeny and phylogeny (see Denison & Xu, in press, for a review). What factors influence children's use of heuristics in probabilistic decision making tasks and what are the most effective methods for teaching children to reason proportionally?

In the present paper we intend to investigate the effect of strategy-specific feedback on children's use of heuristics when reasoning about the outcome of a future event. The US Common Core State Standards recommend introducing children to the formal principles of probability theory in school around the age of 12 (Best Practices, 2017), yet research in developmental psychology suggests that young children and even infants demonstrate accurate intuitions about uncertain outcomes (Denison & Xu, 2014; Falk, Yudilevich-Assouline, & Elstein, 2012; O'Grady & Xu, 2018; Teglas, Girotto, Gonzalez, & Bonatti, 2007; Xu & Garcia, 2008). While many previous studies have sought to observe and explain the features and contexts which influence children's decision making, the current series of studies expands on this work to identify methods for improving these decision-making abilities by reducing children's reliance on inaccurate heuristic decision rules.

4.1.1 Development of probabilistic reasoning

Researchers have studied children's probabilistic reasoning for decades using a simple, 2-alternative forced-choice (2AFC) random draw task in which children are asked to choose between two groups of marbles containing varying amounts of different color marbles with the

goal of drawing a single marble of a target color. Piaget and Inhelder (1975) first developed the random draw task to assess children's ability to use quantity information when making probability judgments. Decades of research on the topic have led to methodological and procedural refinements (Chapman, 1975; Falk, Falk, & Levin, 1980; Falk et al., 2012; Fischbein, Pampu, & Mânzat, 1970; Hoemann & Ross, 1971; Yost, Siegel, & Andrews, 1962). Recently, Falk et al. (2012) devised a strategy assessment task involving 24 trials of the random draw task in order to study children's use of heuristic decision rules.

Falk et al. (2012) identified the four most common strategies that children use in the random draw task and their data suggest that children transition through the four strategies as they learn more about the proportional nature of probability. According to this account, most children begin by focusing on one-dimension of the problem such as the amount of target or non-target events. Children who use these one-dimensional strategies make their decisions by choosing either the group with a greater number of target outcomes ('more favorable') or the group with fewer non-target outcomes ('less unfavorable'). As their learning progresses, they begin to integrate the two-dimensions into more complex decision rules which Falk et al. (2012) label as two-dimensional strategies such as choosing the group with the greater difference between target and non-target marbles ('greater difference') as well as the correct proportional strategy ('greater proportion', i.e., number of winning beads out of the total of winning and losing beads). Results from this study indicate that the use of the 'greater proportion' strategy increases with age from 4 to 11 and that about half of the children begin to reliably use the correct proportional strategy by 8 years of age.

4.1.2 Research on statistics education

Although educators and curriculum development professionals from around the world have called for the advancement of probability literacy in mathematics education, very little attention has been paid to identifying the most effective practices for teaching children about chance (Batanero, Chernoff, Engel, Lee, & Sánchez, 2016). A common them in several reviews of the probabilistic reasoning and mathematics education literatures is a call for educators leverage a student's intuitive understanding of probability to enhance the learning of formal principles (Batanero et al., 2016; Bryant & Nunes, 2012; English & Watson, 2016; Sharma, 2015). How do children develop these intuitive notions about the outcome of future events and how can educators and researchers craft the learning environment in such a way as to loosen children's reliance on inaccurate heuristics and scaffold them toward a more accurate understanding?

In an important review of the literature on statistical education, Garfield and Ahlgren (1988) report that school-age children have difficulty formulating an intuitive understanding of the basics of probability and statistics for several reasons. In their discussion of the factors, Garfield and Ahlgren (1988) point to students' difficulty with rational number and proportions, the conflict between formal probability concepts and students' real-world experience, as well as students' aversion to statistics and probability resulting from learning these concepts at a very abstract and formal level. Although there is still a great deal of research needed on all three of these issues, previous intervention research has addressed several of these concerns.

Previous intervention research has attempted to teach children appropriate strategies for both calculating and reasoning about probability. Fischbein and Gazit (1984) investigated the

influence of computational and conceptual lessons on 10- to 13-year-olds' ability to reason about the results of rolling two dice. Children in the intervention group received 12 lessons on formal probability concepts as well as the use of computational strategies for solving probability problems while children in a control group received their usual mathematics lessons. Results revealed that although the intervention group outperformed the control group on computational questions (i.e. questions that can be solved by applying a specific algorithm), the two groups of students did not differ in performance on conceptual questions (i.e. questions in which children needed to generalize a concept to a novel context). Although Fischbein and Gazit (1984) report the use of a successful intervention, children in the experimental group may have simply learned the computational algorithms needed for solving specific problems without improving their conceptual understand of probability.

Castro (1998) used a didactic approach to teach young high-school students (14-15 year olds) the formal properties of probability. In using this method, teachers in the experimental group were encouraged to probe children's intuitive understanding of probability by posing probability problems and then asking children to explain their own solutions. Students then carried out random experiments to test these ideas and their teachers provided formal explanations for experimental results. Next the students worked with teachers to apply the newly learned concepts to novel contexts and discuss how they might revise their initial ideas based on the results gained throughout the entire learning process. Children in the control group simply received the standard lessons on probability from their mathematics curriculum. Results revealed that students in the experimental group outperformed students in the control group on both probability reasoning and probability calculation post-tests. Importantly, Castro (1998) found that students in the experimental group improved their performance from pre-test scores on both conceptual (i.e. problems requiring the generalization of probability concepts) and computational post-tests (i.e. problems requiring a specific algorithm to compute an answer). Although these findings are exciting, it is important to note that this study included older teens who have already gained a great deal of experience with formal probability concepts. Thus, it is difficult to confirm whether the conceptual change resulted from the specific teaching method or the interaction of instruction and prior knowledge.

In a more recent study, Nunes, Bryant, Evans, Gottardis, and Terlektsi (2014) investigated the influence of additional instruction on 10-year-old children's understanding the role of sample space in calculating probability. In the 'sample space' intervention condition, students received seven, 50-minute lessons in which they learned how to classify sets of outcomes and quantify probabilities based on ratios. In a 'problem solving' control condition children received lessons on mathematical problem solving meant to control for the additional cognitive demands face by the children in the 'sample space' condition without providing information related to the concept of sample space. For example, children in both the 'sample space' and 'problem solving' conditions were taught how to use tree diagrams but children in the 'sample space' condition used this method to solve problems related to the sample space while children in the 'problem solving' condition were given problems not related to the concept of sample space. Finally, children in a second a control condition received their regularly scheduled lessons from their teachers. Children in all three conditions were given a pre-test at the start of the study, as well as several post-tests during and after completion of the teaching phase of the study in order to measure progress on specific lessons and concepts. Results revealed that the 'sample space' intervention group outperformed both control groups on all post-test measures including a delayed post-test presented 2 months after the completion of the experimental

lessons. These findings indicate that when children are provided with lessons in which problem solving solutions are applied to specific probability concepts, they perform better than children who receive traditional mathematics instruction as well as children are taught how to use the same methods to solve mathematics problems more generally.

4.1.3 Prior knowledge and instructional context

Socio-cultural approaches to cognitive development and education argue that teachers and their students can form different assumptions about communicative exchanges which can be problematic during instruction if both teacher and learner believe the same mathematical function (i.e. calculating probability) is supported by different forms (i.e. absolute number, ratio, or proportion). Mathematics educators have long understood the power of identifying and leveraging a student's intuitions and prior conceptual knowledge in scaffolding the learner to a more thorough understanding of mathematical concepts. As an example, Saxe, Gearhart, and Seltzer (1999) investigated the interaction between children's prior knowledge of fractions and the practices their teachers employed in the mathematics classroom. Children were assessed as either having or not having a rudimentary part-whole understanding of fractions while their teachers were rated on a scale representing the degree to which their teaching methods aligned with reform standards. Importantly, the researchers characterized, 'high alignment' as the degree to which a teacher engages with a student's existing mathematical knowledge as well as their engagement of conceptual issues during problem solving tasks. Conversely, classrooms that focus either on self-discovery or procedural memorization were considered low in alignment with reform policies. Findings indicated that high classroom alignment predicted improved performance on a post-test and this effect was greater for children without a rudimentary part-whole understanding of fractions. With low levels of alignment to reform principles, students without a rudimentary understanding likely relied on their prior conceptual understanding of integers. However, with supportive classroom environments in which teachers seek to identify and build upon a student's prior conceptual understanding, children are in a better position to construct a thorough and accurate understanding. These findings indicate that the instructional context as well as a learner's prior conceptual understanding interact to influence learning outcomes. Teachers who are capable of identifying a child's prior conceptual knowledge of fractions can construct learning environments that either confirm accurate conceptualization or scaffold learners toward a more thorough understanding.

4.1.4 Rationale

While it is clear that children's strategy use improves with age and education it is unclear how this learning process unfolds. In the current series of studies, we investigate whether children's use of heuristic decision rules is influenced by feedback. In Experiment 1, we investigate the influence of feedback on children's strategy use by first assessing children's strategies and then presenting them with a series of 2AFC random draw task trials during which children are given feedback. Finally, we present test trials designed to investigate whether children changed their strategy. Based on previous research (Falk et al., 2012), we hypothesize that young children are capable of learning to use the correct strategy but only when provided with examples that do not fit their incorrect understanding (i.e. children must receive feedback on trials which conflict with their strategy in order to learn the proportional nature of simple

probability). We predict that children can learn to make correct choices in the 2AFC random draw task if they are presented with trials that conflict with their strategy.

4.2 Experiment 1

4.2.1 Methods

Participants. Fifty-seven children between the ages of 6 and 11 were recruited from museums, schools, and homes in the San Francisco Bay area. Data from ten children were excluded from this sample: One child decided to stop the game early, three children were coached by their parents, and four children were excluded due to equipment malfunction or experimenter error. An additional two children were excluded because their average reaction time on the ***assessment phase*** was lower than 3 seconds and thus did not have enough time to count the marbles as instructed. Our final sample consisted of N = 47 children (5 6-year-olds, Mean age = 6.52, SD = 0.13; 5 7-year-olds, Mean age = 7.50, SD = 0.29; 7 8-year-olds, Mean age = 8.53, SD = 0.25; 17 9-year-olds, Mean age = 9.50, SD = 0.32; 9 10-year-olds, Mean age = 10.50, SD = 0.29;4 11-year-olds, Mean age = 11.64, SD = 0.34).

Material. Images depicting two gumball machines and two groups of green and purple marbles were rendered using Blender (Version 2.78) 3D animation software. Following Falk et al. (2012), each trial image was internally labeled with the trial type designators 'GGGG', 'GGGS', 'SSSG', and 'SSSS' with each letter representing the dimension of comparison and the letter itself relating the correct choice (higher probability of yielding the child's favored color marble) to the incorrect choice (lower probability of yielding the child's favored color). For each target color (i.e. green or purple), two sets of 24 images were created using the same distributions used by Falk et al. (2012) for a total of 96 images.

Figure 1 presents an example image for each trial type. Note that the correct choice in the figure on the top left (labeled 'GGGG') has a greater amount of favored marbles (1st G), a greater amount of non-favored marbles (2nd G), a greater total of favored and non-favored marbles (3rd G) and a greater difference between favored and non-favored marbles (4th G). In contrast, the correct choice for the image on the top right (labeled 'SSSS') has a smaller amount of marbles in each of these categories compared to the incorrect choice. Children using a strict 'more favorable' strategy would make a correct choice on all 12 'GGGG' and 'GGGS' trials but would choose incorrectly on all 12 'SSSS' and 'SSSG' trials. A child using a strict 'less unfavorable' strategy would make a correct choice on all 12 'SSSS' and 'SSSG' trials but would choose incorrectly on all 12 'GGGG' and 'GGGS' trials. Children using a strict, 'greater difference' would make a correct choice on all 12 'GGGG' and 'SSSG' trials but would choose incorrectly on all 12 'SSSS' and 'GGGS' trials. Finally, a child using the formally correct proportional strategy would choose correctly on all 24 trials.

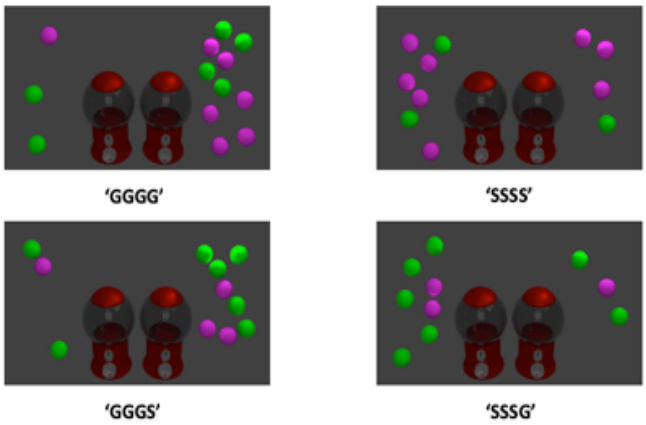

Figure 4.1 Example images for each of the 4 trial types used in Experiment 2. In all 4 images, the correct choice for obtaining a purple marble is located on the right side of the image

Procedure. Experiment 2 consisted of three phases. First, children performed the 'exact-numerical-value- probability task' in the ***assessment phase*** in order to identify the child's strategy. The Matlab program recorded the participant's choices and determined the participant's strategy score. Children using the 'more favorable' strategy were coded as '1', 'less unfavorable' strategy was coded as '2', 'greater difference' strategy was coded as '3' and the correct proportional strategy was coded as '4'.

In the ***conflict phase***, children were semi-randomly assigned to one of two conditions in which they were given feedback about their choices. For each group of strategy users, children were assigned evenly and pseudo-randomly into 'high conflict' or 'low conflict' conditions consisting of 12 trials. We chose this method to ensure that there were an equal number of children using each strategy in both high and low conflict conditions. Children were told that in this part of the game they will get to see what color marble they get by looking in the tray of the gumball machine that they chose. Since there was no effect of target color in either task of Experiment 1 we decided that all participants would be asked to collect green marbles. Importantly, feedback was given deterministically, meaning that if the child made the mathematically correct choice, they receive a green marble and if they chose incorrectly they received a purple marble. The set of 12 conflict trials were matched to the strategies children used such that if the child used their strategy on every trial they would receive 12 purple marbles (instead of the 12 green marbles they were trying to get) and thus children in this condition experience higher conflict between the predictions of their strategy and the actual outcomes. In contrast, children in the low conflict condition as well as children who used the correct proportional strategy during the ***assessment phase*** were given 12 trials randomly selected from the set of 24 trials. Due to the random trial presentation, some trials in the low conflict condition will conflict with their strategy and provide negative feedback while other trials are in agreement with their strategy and provided positive feedback. The high conflict condition represents a

guided learning scenario in which the teacher (in this case the Matlab program) knows the child's level of understanding and provides the type of examples necessary for the child to overcome their errors. In contrast, the low conflict condition provides a baseline since the feedback trials were chosen randomly.

Finally, during the ***test phase***, the children were asked to play 4 more trials in which they can win prizes depending on how many green marbles they get. Before the beginning of the test phase, children are reminded that they should count the number of marbles in all of the groups and that they can take as long as they need to make a decision. Children's responses were recorded and all participants received 2 prizes to thank them for participating regardless of the number of green marbles they collected. In this preliminary task we decided to present children with 4 test trials rather than the full set of trials used in the ***assessment phase*** in order to keep the overall time for the experiment below 20 minutes in length.

4.2.2 Results

Results from the ***assessment phase*** indicated that the majority of children in Experiment 1 utilized one-dimensional strategies. 27 children (57.45%) used the 'more favorable' strategy, 5 children (10.64%) used the 'less unfavorable' strategy, 8 children (17.02%) used the 'greater difference' strategy, and 7 children (14.89%) used the correct proportional strategy. Figure 2 presents the proportion of children using each strategy.

In order to compare children in high and low conflict conditions we calculated the average number of correct responses for each child in both the ***assessment phase*** and ***test phase***. Importantly, children who used the correct proportional strategy were not included in the analyses of the ***conflict*** and ***test phases*** because they could not be assigned to a high conflict condition. For the ***assessment phase***, children in the high conflict condition were not significantly different from those in the low conflict condition ($\Delta M = 0.03$, 95% CI $[-0.13, 0.07]$, $t(35.50) = -0.62$, $p = .537$). However, during the ***test phase***, children in the high conflict condition (74% correct) performed significantly better than children in the low conflict condition (32% correct; $\Delta M = -0.42$, 95% CI $[0.26, 0.58]$, $t(36.63) = 5.29$, $p < .001$). Figure 3 presents the average performance of children in both conflict conditions.

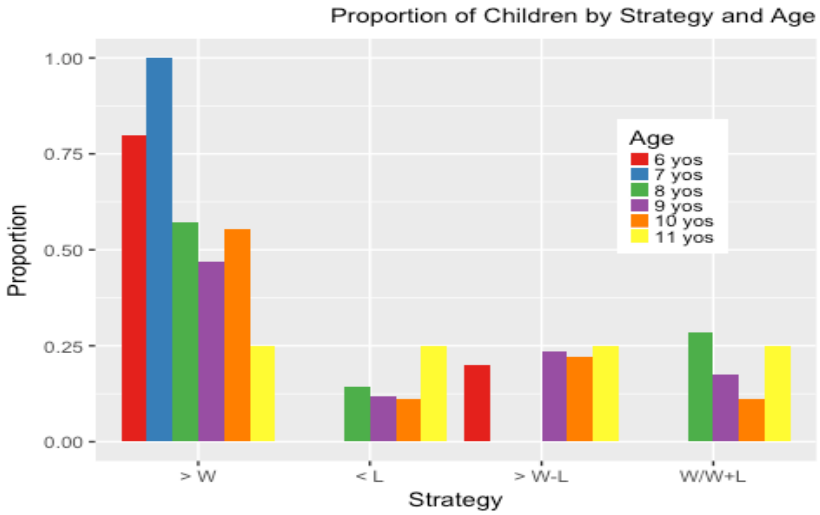

Figure 4.2 Proportion of children using each strategy by age group.

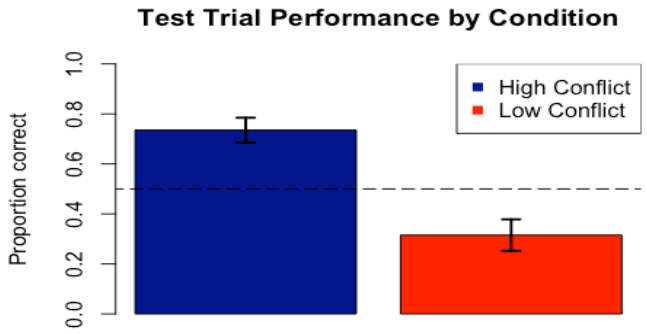

Figure 4.3 Average test trial performance by condition.

4.2.3 Discussion

Results from the ***conflict*** and ***test phases*** indicated that children were able to switch strategies after being provided with enough negative feedback using trials which conflicted with their strategy suggesting that younger children are capable of using the correct proportional strategy if they are provided with enough evidence that their original strategy is not working.

However, the ***test phase*** in Experiment 1 cannot determine whether children actually adopted a different strategy or simply learned to 'choose the opposite' of their inaccurate strategy. For this reason, we designed Experiment 2 with two important innovations. First, we included a full post-test phase consisting of the same 24 trials included in the ***assessment phase*** in order to compare patterns of performance across the various trial types. Second, we asked children to provide verbal reports of their strategy use after each phase of the experiment. If children are simply learning to 'pick the opposite', they will have an opposite response pattern in their post-test data which should be confirmed by verbal reports of using the 'pick the opposite' strategy.

4.3 Experiment 2

4.3.1 Methods

Participants. The current experiment was pre-registered (http://aspredicted.org/blind.php?x=mp6gc9) with a target sample of 80 children between the ages of 7 and 10 (20 children in each age group: 7-year-olds, 8-year-olds, 9-year-olds, and 10-year-olds), which was determined based on previous research using a similar task (Falk et al., 2012). Data collection continued from June 2018 to June 2019 in order to prevent overlapping between the two age groups. For example, 7-year-olds recruited at the beginning of the study would have aged into the 8-year-old group by the conclusion of the study. For this reason, we had to stop data collection at N = 72 children (20 7-year-olds, Mean age = 7.43, SD = 0.26; 19 8-year-olds, Mean age = 8.19, SD = 0.27; 17 9-year-olds, Mean age = 9.13, SD = 0.25; and 16 10-year-olds, Mean age = 10.11, SD = 0.34). All 72 children participated in the first session and eight children declined to participate in the follow-up session 1 week later (2 7-year-olds, 3 8-year-olds, 2 9-year-olds, and 1 10-year-old).

Material. We used the same procedure for creating images as reported in Experiment 1.

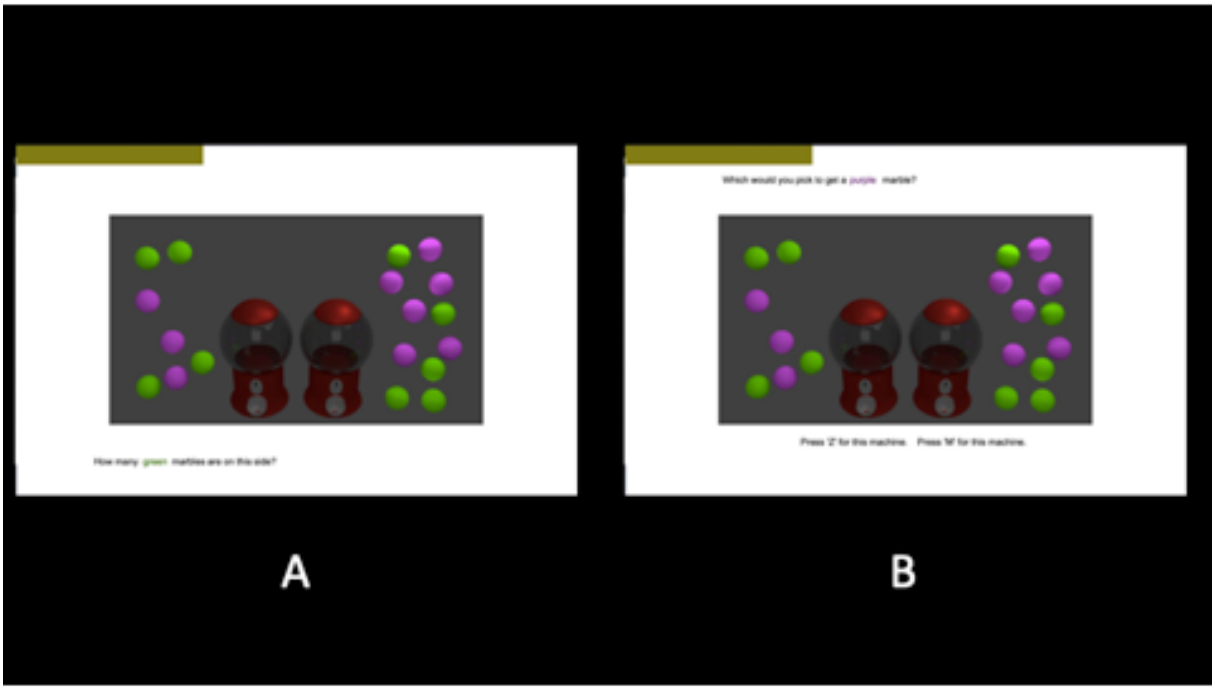

Figure 4.4 Two screen shots of a game in progress during the assessment phase. A) Example of a counting prompt. B) Example of a choice prompt.

Procedure. Children were seated approximately 60 cm away from a MacBook Pro laptop (OSX; Screen resolution 1280 x 800) and told they would play a game in which they would try to collect green or purple marbles from one of two different gumball machines. The task consisted of a self-paced game automated using the psychophysics toolbox written for the MatLab programming language (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997). In order to maintain an average testing time of 20 minutes to prevent fatigue, the experiment was split into 2 testing sessions spaced 1 week apart. Children completed the *assessment phase* during session 1 and then completed the *conflict phase* and *post-test phase* during session 2.

Assessment Phase. During the first testing session, the experimenter explained the task and children were prompted to choose their favorite of the two colors, either purple or green. For each of the 24 images, the computer presented the image at random along with 4 counting prompts, one for each group of marbles (i.e. "How many (green/purple) marbles are on this side (left/right)?"). The child responded by pressing the appropriate number key on the keyboard. An error message was presented if the child chose the wrong number and the game did not progress until the child pressed the correct number key. Counting prompts for each color and side were

randomized for each image. Once children completed the counting prompts, they were prompted with the question "Which would you pick to get a (green/purple) marble?". Importantly, the position of the marbles on the screen were randomized to prevent children from choosing based on the positions of their favorite color marble. However, in order to ensure that children did not rely on the placement of the marbles on the screen they were told to "Pretend that the marbles will go into the machines and that the machines will be shaken up so you don't know what's going to come out next." After completing all 24 trials, children were asked to provide a verbal report of the strategy. To generate these verbal reports, the experimenter asked, 'Can you tell me what helped you make your choices in this game?'. If the children replied 'I don't know' or provided an irrelevant answer (i.e. 'I like the purple ones') the experimenter then provided appropriate follow-up questions such as 'Some kids who have played this game say that they look at the two groups and pick whichever has more [purple] because they like those, but other kids say that they pick the one that has the least [green] because that's the one they don't want. Did you think about these things or did you pick based on something else?'.

Following the methods outlined by Falk et al. (2012), the MatLab program discerned which strategy the child used based on their performance on each of the four trial types. After children completed the *assessment phase*, the MatLab program calculated point scores for each strategy based on the choices that the child made. Whichever strategy had the highest point score was deemed to be that child's strategy. Point scores could range from 0 to 24 with 0 indicating no strategy-consistent responses and 24 indicating perfect strategy use. Participants using the 'more favorable' strategy provided about 21 (M = 21.79; SD = 3.15) out of 24 strategy-consistent responses, while those using the 'less unfavorable' strategy provided 16 (M = 15.71; SD = 1.80) out of 24 strategy-consistent responses, participants using the 'greater difference' strategy provided 20 (M = 20.40; SD = 20.40) out of 24 strategy-consistent responses, and participants using the 'greater proportion' strategy provided 19 (M = 20.31; SD = 20.31) out of 24 strategy-consistent responses.

Conflict Phase. Children were semi-randomly assigned to one of two different conditions ensuring that an equal number of children using each strategy were assigned to both conditions. In the half-conflict condition, children viewed all 24 trials, 12 of which conflicted with the child's strategy and 12 of which did not conflict. Children in the high-conflict condition viewed 24 trials that conflicted with their strategy.

In the 'high conflict' condition, feedback trials were assigned as follows. Children designated as using the 'more favorable' strategy viewed 12 'SSSS' trials and 12 'SSSG' trials. Children using the 'less favorable' strategy viewed 12 'GGGG' trials and 12 'GGGS' trials. A child using the 'greater difference' strategy viewed 12 'SSSS' trials and 12 'GGGS' trials while children using the proportional strategy were simply assigned to the half-conflict condition as none of the trials conflicted with their strategy. In all conditions and for all trials, children received feedback in the form of either a favored or unfavored color marble returned in the dispenser of the machine they chose. Importantly, all feedback was provided deterministically, meaning that if a child chose strictly according to their non-proportional strategy in the high-conflict condition, they would receive 24 unfavored marbles and a child in the half-conflict condition would receive 12 favored and 12 unfavored marbles. Children in the half-conflict condition received a mix of confirmatory and dis-confirmatory feedback with respect to their strategy while children in the high-conflict condition received only dis-confirmatory feedback.

Post-test Phase. After completing the *conflict phase*, each child received the same 24 trials they viewed in the *assessment phase* one week prior in a randomized order. Importantly,

the *post-test phase* is an immediate post-test because it occurred directly following the *conflict phase*. After completing the *post-test phase* the experimenter prompted the child to provide a verbal report of their strategy similar to that provided after the assessment phase. In addition, the experimenter asked a follow-up question, 'Did you change your strategy based on the color of marbles you received during the first part of today's game (*conflict phase*)?'.

Analysis and coding. Children's verbal reports of their strategy use in the *assessment phase* were coded by a research assistant who was blind to the study hypothesis and conditions. The same research assistant also coded children's responses to the strategy change follow-up question posed after completion of the *post-test phase*.

The results of each phase of the experiment are reported separately below along with a brief discussion section. For all three phases, analyses consisted of comparisons of Generalized Linear Regression Models with Mixed effects (GLMMs) using the lme4 package written for the R statistical programming language (Bates, Maechler, Bolker, & Walker, 2015). All models predicted the binary response variable while holding participant ID as a random effect. Nested models were compared using Chi Squared tests for model fits while non-nested models were compared using the Akaike Information Criterion (AIC), a measure of model fit in which models with smaller AICs are preferred over models with higher AICs. For all three phases of the experiment, modeling results revealed no influence of participant gender, favored color, on performance. Model coefficients for GLMMs are reported as log-odds, that is, the log of the odds ratio of correct to incorrect responses.

4.3.2 Results

Assessment Phase Results. In order to investigate the degree to which children explicitly use a specific strategy in the current task we calculated inter-rater reliability between the children's strategies as derived by the computer based on their performance across the four trial types and the strategies inferred by the independent coding of children's verbal responses after the assessment phase. Results revealed moderate agreement between the independent coder and the children's strategies (Cohen's $\kappa = 0.56$, $p < .001$). Interestingly, Cohen's κ values were higher for children who used the simple, one-dimensional 'more favorable' and 'less unfavorable' strategies (Cohen's $\kappa = 0.70$, $p < .001$) likely reflecting the difficulty children have when verbally expressing more complex two dimensional strategies such as 'greater difference' and 'greater proportion'.

Comparisons of GLMMs revealed that the model with the best fit to the assessment phase data was the model predicting performance from strategy alone ($AIC_{Strategy}$ =2,226.25). This model outperformed the null model (AIC_{null} = 2,311.07; χ^2= 90.82; df = 3; $p < .001$), as well as the model predicting performance from age (AIC_{Age} =2,303.64). More complex models predicting performance from age and strategy ($AIC_{Strat+Age}$ = 2,225.88; χ^2= 2.37; df = 1; p = .12) and the interaction of age and strategy ($AIC_{Strat*Age}$ = 2,227.81; χ^2= 6.44; df = 4; p = .17) did not perform better than the model for strategy alone. Thus, the simpler model is preferred since it can predict the same amount of variance with fewer model parameters. There was no effect of trial number indicating that children's performance did not improve with time during the *assessment phase*.

Inspection of model coefficients reveals that the log-odds of a correct response increased for children using the 'greater difference' ($\beta_{>F-U}$ = 0.38; SE = 0.15; 95% CI [0.09, 0.67]), and 'less unfavorable' strategies ($\beta_{<U}$ = 0.18; SE = 0.17; 95% CI [-0.15, 0.51]), as well as those using formally correct proportional strategy ($\beta_{>F/F+U}$ = 1.65; SE = 0.17; 95% CI [1.31, 1.98]) compared to children using the 'more favorable' ($\beta_{>F(Intercept)} = 4.003^{-16}$; SE = 0.06; 95% CI [-0.06, 0.18]). However, only the coefficients for 'greater difference' and 'greater proportion' strategies reached statistical significance (Wald test: 'greater difference': $p < .01$; 'greater proportion': $p < .001$). Figure 2 presents the proportions of children using each strategy by age. Note that the two younger age groups (7-year-olds and 8-year-olds) are predominantly relying on the 'more favorable' strategy whereas children in the two older age groups (9-year-olds and 10-year-olds) have a more equal spread across the four different strategies.

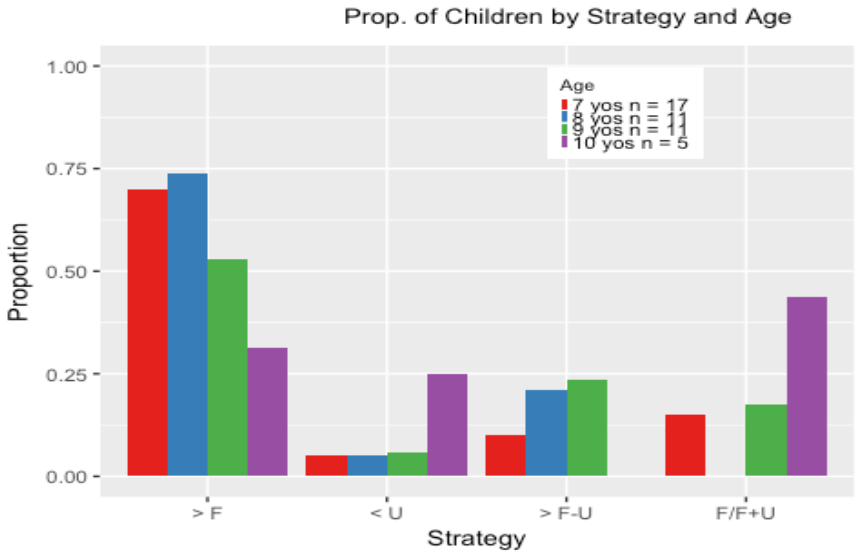

Figure 4.5 Proportion of children using each strategy by age group. Strategies are designated as follows: '> F': 'more favorable'; '< U': less unfavorable'; '> F-U': 'greater difference'; '> F/F+U': 'greater proportion'.

Assessment Phase Discussion. Results of the current study converge with those of previous reports indicating that children's use of the correct proportional strategy improves with age (Falk et al., 2012; O'Grady & Xu, 2019). Importantly, results of the GLMM comparisons revealed an effect of strategy on performance indicating that children who attended to the number of both favorable and unfavorable marbles in each choice performed better than children who made their choices based on one single dimension (i.e. choosing based solely on the number of favorable or unfavorable marbles).

Conflict Phase Results. Thirteen children were found to be using the formally correct proportional strategy during the assessment phase the previous week. Since there are no trials that conflict with this strategy, data from these children were excluded from the conflict phase analyses. In addition, the 8 children who did not return for session 2 did not contribute data to either the conflict or post-test phases resulting in a sample size of N = 51. Comparisons of GLMMs for this subsample revealed that the model with the best fit to the data predicted performance from the conflict condition and the trial number as well as the interaction between

the two variables ($AIC_{Condition*Trial}$ =1,504.84). This model outperformed the null model (AIC_{null} = 2,311.07; χ^2= 38.75; *df* = 3; $p < .001$) as well as the simpler models predicting performance from conflict condition ($AIC_{Condition}$ = 1,532.28; χ^2= 31.44; *df* = 2; $p < .001$) and trial number alone (AIC_{Trial} = 1,528.74; χ^2= 27.91; *df* = 2; $p < .001$) and the more complex model accounting for both the conflict condition and trial number without an interaction ($AIC_{Condition+Trial}$ = 1,523.45; χ^2= 20.61; *df* = 1; $p < .001$). There were no significant effects of the three non-proportional strategies, nor was there an interaction between conflict condition.

Inspection of the model coefficients revealed that the log-odds of a correct response decreased for children in the high-conflict condition ($\beta_{High-Conflict}$ = -0.30; SE = 0.35; 95% CI [-0.98, 0.38]) compared to children in the half-conflict condition ($\beta_{Half-Conflict}$ = 0.31; SE = 0.24; 95% CI [-0.16, 0.79]), which is not surprising considering that all of the trials in the high-conflict condition conflicted with the children's strategies whereas only 12 of the 24 trials in the half-conflict condition conflicted with the children's strategies. While the model coefficient for trial number was slightly negative (β_{Trial} = -0.01; SE = 0.01; 95% CI [-0.03, 0.02]) indicating a decrease in the log-odds of a correct response, the interaction between trial number and condition revealed that in the high-conflict condition, trial number had a positive effect on the log-odds ($\beta_{Trial*High-Conflict}$ = 0.09; SE = 0.02; 95% CI [0.05, 0.12]). The interaction between trial order and the high-conflict condition was the only model coefficient to reach statistical significance (Wald test: $p < .001$) indicating that performance improved over time in the high-conflict condition suggesting that children in this condition may have learned from feedback on earlier trials.

Conflict Phase Discussion. Results revealed that both conflict condition and trial order had an effect on performance. Importantly, the interaction between conflict condition and trial number produced the greatest positive effect on performance while the coefficient for the high-conflict condition alone had a negative effect on performance. This set of results suggests that children in the high-conflict condition began by choosing according to their strategy but then switched to another strategy after several trials in which they received negative feedback.

Post-Test Phase Results. Of the 13 children who used the correct proportional strategy during the *assessment phase* only one child (a 10-year-old) did not continue to use the correct proportional strategy. Interestingly, this child used the 'more favorable' strategy and reported that they switched to a simpler strategy because "the game was boring and I wanted to finish it faster" suggesting that this child understood the time-accuracy tradeoff among the various potential strategies. Table 1 presents the number of children using each strategy in the *post-test phase* ('Post-test' column) based on the child's *assessment phase* strategy ('Assessment' column) and condition ('half-conflict' and 'high-conflict' columns).

Comparisons of GLMMs revealed that the model with the best fit to the data predicted *post-test phase* performance based on the interaction between conflict condition and *assessment phase* strategy ($AIC_{Condition*Strategy}$ =1,613.06). This model outperformed the null model (AIC_{null} = 2,311.07; χ^2= 24.82; *df* = 5; $p < .001$), the models for conflict condition alone ($AIC_{Condition}$ = 1,623.55; χ^2= 18.49; *df* = 4; $p < .001$) and *assessment phase* strategy alone ($AIC_{Strategy}$ = 1,623.70; χ^2= 16.64; *df* = 3; $p = .00$) as well as the model for conflict condition and *assessment phase* without any interactions ($AIC_{Condition+Strategy}$ = 1,617.89; χ^2= 8.84; *df* = 2; $p = .01$).

Inspection of model coefficients revealed that the only model coefficient to reach statistical significance was the interaction between 'greater difference' strategy and the high-conflict condition which increased the log-odds of a correct response ($\beta_{\prime>F-U\prime*High-Conflict}$ = 1.32; SE = 0.45; 95% CI [0.45, 2.20]; Wald test: $p < 0.01$) compared to the children using the 'more favorable' strategy in the half-conflict condition ($\beta_{Intercept}$ = -0.02; SE = 0.26; 95% CI [-0.53, 0.49]). All remaining coefficients did not reach statistical significance. Figure 3 presents the proportion of correct responses by conflict condition and *assessment phase* strategy.

Verbal report data. Analyses of children's' verbal reports of strategy change after the post-test phase revealed that 4 of the 29 children who changed strategies (~13.8%) reported using a 'pick the opposite' strategy. For example, when asked 'Did you ever change strategies as a result of the marbles that you received during the first part of today's game?' one child whose target color was green marbles said, "...if you pick the one with the greatest number of green marbles, it will probably give you a purple". Another 4 out of 29 (~13.8%) children claimed to have used the same strategy throughout the study even though the computer recorded these children as using a different strategy during the post-test phase. Finally, the verbal reports of most of the remaining children recorded as changing strategies indicated that they believed they used the correct proportional strategy (11 out of 29 or ~37.9%), while only 2 out of 29 (~6.9%) children reported using the 'more favorable' strategy, 3 out of 29 (10.3%) children reported using the 'greater difference' strategy during the post-test, and 4 out of 29 (~13.8%) children reported using the 'less unfavorable' strategy. One child recorded as changing strategies indicated that they did not know if they changed strategies.

Post-Test Phase Discussion. The interaction between conflict condition and *assessment phase* strategy indicates that children using the 'greater difference' strategy benefited more from the high-conflict condition compared to children using the other 2 strategies ('more favorable' and 'less unfavorable'). Although these findings are promising, there were only three children using the 'greater difference' and three children using the 'less unfavorable' assigned to the high-conflict condition thus more data will be needed to make any firm conclusions. However, it is interesting to view the differences between the 'half-conflict' and high-conflict conditions for children using the 'more favorable' strategy. Note from Table 1 that only 3 of the 15 children using this strategy in the high-conflict condition (20%) continued using their strategy after the feedback condition while 9 of the 13 assigned to the half-conflict condition (69.2%) continued to use the strategy.

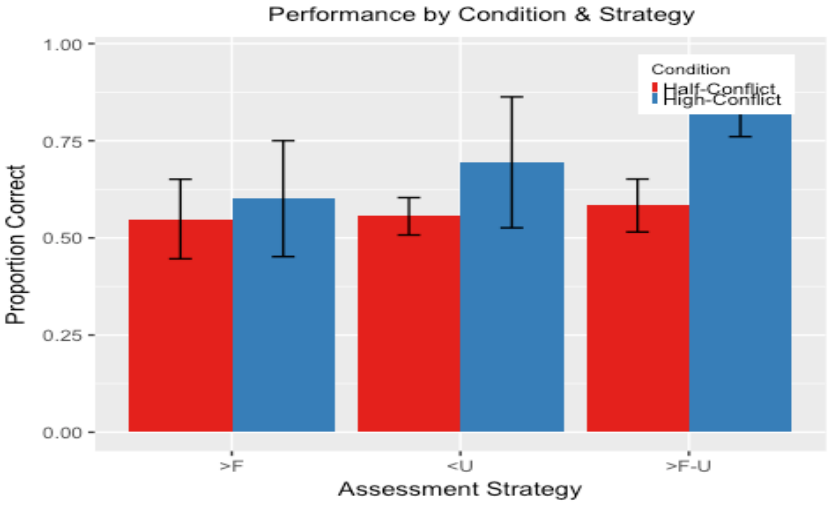

Figure 4.6 Proportion of correct responses in the post-test phase by conflict condition and assessment phase strategy. Error bars indicate standard deviation.

4.4 General Discussion

In Experiments 1 and 2 we find that children in the high conflict conditions were more likely to change their decision making strategy when they are provided with consistent feedback indicating that their heuristic decision rule is inaccurate. These findings indicate that children can override their reliance on inaccurate heuristic decision rules if they are given enough strategy-specific feedback. Importantly, children in all conditions across both experiments received negative feedback on their incorrect strategy, however, since children in the low conflict condition of Experiment 1 and half conflict of Experiment 2 receive a mix of conflicting and non-conflicting trails they receive both positive and negative feedback. While more work is needed to uncover the influence of this mixed feedback it is interesting to note that even Piaget credited children at this age with the understanding of the uncertain nature of chance (Piaget & Inhelder, 1955). Our data suggest that this prior understanding of uncertainty interferes with children's ability to learn from negative feedback in the low and half conflict conditions. Simply put, children in these conditions just shrug off the fact that their strategy did not yield the target outcome as mere chance and they see the fact that the receive at least some of their favorite marbles as confirmation of this notion. Future work should address this potential interpretation by assessing children's knowledge of uncertain outcomes before providing feedback.

Children do not enter the classroom as a blank slate, rather they carry with them their intuitions and prior knowledge derived from their experience. Modern constructivist theories use the language of Bayesian probability to express developmental change as the integration of prior beliefs and new data (Fedyk & Xu, 2018; Gopnik & Wellman, 2012; Xu, in press). One of the most important responsibilities of a teacher is to identify a learner current knowledge state in order to design effective instruction. This approach is critical when teaching probability because the inherent uncertainty adds noise to the learning signal. In Experiments 1 and 2 we demonstrate that children abandon their probabilistic decision making strategy when given examples that conflict with their prior beliefs (high conflict condition) but they tend to stick with their inaccurate strategies when provided with a mix of both confirming and disconfirming examples

(low conflict). These results indicate that a learner's prior conceptualization of probability influences how they respond to feedback.

While exciting and illuminating, the current series of experiments has some important limitations. First, since Experiment 2 employed an immediate post-test, it is unclear whether children's responses reflect bona fide conceptual change or a merely temporary shift of attention to different stimlulus features. Follow-up research involving a delayed post-test will be needed to ensure that children fully incorporated their new understanding of probabilty rather than temporarily shifting their decision-making strategy.

Another important limitation was partially addressed by the full post-test and verbal reports recorded in Experiment 2: do children really switch strategies or just learn to 'pick the opposite'?. Although the post-test assessment employed in Experiment 1 could not adequately identify which strategy children were switching to, the improved post-test assessment and qualitative data from children's' verbal reports in Experiment 2 indicated that children understood that their initial strategy was incorrect and that they needed to formulate a different decision rule. While the post-test and verbal reports in Experiment 2 were meant to address this concern, the quality of children's verbal reports may not give a full and clear picture of the extent to which conceptual change has occurred. Future studies will involve a longitudinal design in which children and families are invited back into the lab and provided with feedback on their post-test strategy. If children are merely using a 'pick the opposite' strategy, then an additional round of feedback would cause them to revert back to their original strategy. However, if children are truly engaging in conceptual change, these children should begin to use more complex, two dimensional strategies such as 'greater difference' and 'greater proportion'. Interestingly, these findings would also give weight to Falk et al. (2012)'s claim that children traverse the four strategies ('greater favorable', 'less unfavorable', 'greater difference', and 'greater proportion') as they learn to attend to more complicated features.

One interesting question this work leaves open for future research is the performance gap between infants and young school-age children. As mentioned previously, infants (Xu & Garcia, 2008) and toddlers (Denison & Xu, 2014) appear to be equipped with an intuitive sense of the proportional nature of probability. Denison and Xu (2014) presented 12-month-old infants with a simplified version of the 2AFC random draw task involving two groups of pink and black lollipops. An experimenter randomly selected a lollipop from each group and then placed them in opaque containers. The infant was then allowed to crawl toward the container of their choice to retrieve their prize. Using this task, Denison and Xu (2014) showed that infants crawl toward the group with the greater proportion of favored objects (i.e. for most infants, the pink lollipops) even when presented with a choice between a smaller proportion with a greater absolute number of pink lollipops and a larger proportion with fewer pink lollipops. There are several reasons why infants may demonstrate an intuitive understanding of the proportional nature of probability but older, school-age children rely on heuristic decision rules. One possibility is that infants may have an implicit understanding of probability (see similar claims for the development of theory of mind and object cognition, e.g., Carey, 2009, for a review), and older children's more mature understanding is either quite distinct from the infants' or may be the result of reconstructing these earlier concepts. Either of these developmental processes may follow an extended trajectory and require genuine conceptual change. Future work will hopefully uncover the origin of the heuristic bias in children by investigating the critical transition from early toddler hood to the school-age years.

Children and adults use their knowledge of probability to inform their decision several times a day. While previous research has claimed to show that both children (Piaget & Inhelder, 1975) and adults (Kahneman, 2011; Kahneman & Tversky, 1973; Tversky & Kahneman, 1983) are impoverished decision-makers, more recent evidence suggests that probabilistic reasoning is part and parcel of primate existence (De Petrillo & Rosati, 2019; Denison & Xu, 2014; Rakoczy et al., 2014; Tecwyn et al., 2017). In the current study, we hope to have demonstrated that although children often rely on inaccurate heuristics when reasoning about the probability of future events, their ability to flexibly adapt in response to feedback can help to override these fast yet inaccurate strategies.

Chapter 5

Conclusion

5.1 Conclusions and Implications of Empirical Work

Probabilistic data are ubiquitous in human experience and probabilistic reasoning is a powerful tool for sorting through this variability. How do we represent probability and how do these representations influence our decision making? What is the developmental trajectory of probabilistic reasoning in humans and how do different educational contexts influence children's understanding of the proportional nature of probability? In my dissertation, I have provided evidence to suggest that rapid, non-symbolic probability judgments are surprisingly accurate and rely on approximate representations of number (Chapter 2), the acuity of this ability improves with age but is influenced by incorrect heuristic decision rules (Chapter 3), and these incorrect heuristic decision strategies can be overcome with the proper amount and type of feedback (Chapter 4).

5.1.1 The Psychophysical properties of non-symbolic probability judgments.

In Chapter 2 my collaborators and I developed a 2-alternative forced choice (2AFC) ratio comparison task framed as a probabilistic decision-making experiment. In 4 experiments, we presented undergraduate students at UC Berkeley (Experiment 2.1) and adults recruited online through Amazon's Mechanical Turk (Experiments 2.2, 2.3, and 2.4) with a 2AFC task in which they were instructed to pick one of two binary distributions of colored marbles based on the likelihood of drawing a single marble of a particular color.

Data from Experiment 2.1 indicated that while adult participants demonstrated near ceiling performance on most of the ratios of proportions presented (least difficult: 70% vs 10%; most difficult: 50% vs 55%), they also demonstrated a bias toward choices with larger numbers of target marbles. When participants were presented with trials in which the bin with the smaller proportion of marbles contained a larger absolute number of target marbles compared to the bin with the higher proportion they sometimes chose the lower proportion containing more target marbles. These results indicate that non-symbolic probability judgments are characterized by ratio dependence and are biased toward groups with greater numbers of target events.

Results from Experiment 2.2 replicated these findings and extended the research to account for both numerical and non-numerical stimulus features. Our data suggest that while people can make accurate, rapid judgments about the probability of future events based on proportions, their decisions are often biased toward groups with a larger number of target marbles and groups with larger marbles (i.e. more surface area). Furthermore, when compared to predictions made by a psychophysical model adapted from DeWind et al. (2015) to account for numerical and non-numerical stimulus features in probability judgments, our data suggest that participants compute proportions using approximate number system representations.

In these findings were further replicated and extended in Experiments 2.3 and 2.4. In both experiments I demonstrate that the use of hueristic decision rules by adult decision makers is not

the result of the brief presentation time used in Experiments 2.1 and 2.2. Furthermore, in experiment 2.4, I demonstrate that adult reasoners continue to show a bias on number vs proportion trials even when they are given explicit instructions indicating that this bias is inaccuarate. Together, these findings suggest that while human probabilistic reasoning is quite sophisticated, even educated adults tend to rely on inaccurate heuristics in the face of uncertain data.

5.1.2 The Developmental trajectory of non-symbolic probability judgments

In Chapter 3 I used the same probability tasks deployed with adults in Chapter 2 to demonstrate that school-age children's probability judgments improve with age, are characterized by ratio dependence, share some signature features of the approximate number system, and are biased toward distributions with a greater number of target events. In Experiment 3.1 we manipulate the ratio of the proportions of the compared distributions as well as whether the sample spaces were equal ('total equal' trials) or unequal ('target equal' trials). In Experiment 3.2 we modify this method in order to account for the equality of the sample spaces ('total equal' trials), the number of target events ('number vs proportion' trials), as well as the area of the individual marbles ('area anticorrelated' trials).

Data from Experiment 3.1 suggest that children's probability judgments are characterized by ratio dependence. While 7-year-old children performed better than 6-year-old children, both age groups showed a bias toward choices with a larger number of target events. Results of Experiment 3.2 replicated the main results of Experiment 1 and extended these findings by revealing that children's judgments are influenced by the surface area of the marbles as well as by the number of target events. Regression analyses revealed main effects for age group, trial type, and ratio of proportions as well as interactions between the three variables. Together these results indicate that 8-year-old children are capable of rapidly computing probabilities based on proportions and that the acuity of this rapid approximation improves with age. Furthermore, based on data from the number vs proportion trials, we also find that children's simple probability judgments, much like the adults reported in Chapter 1, are influenced by the number of target marbles indicating that children sometimes rely on incorrect judgmental heuristics.

5.1.3 Feedback influences children's use of heuristic decision rules

Although it seems reasonable to assume that education influenced children's decision-making strategies in Experiments 3.1 and 3.2, simple procedural factors such as presentation time could also explain our results. For example, children were presented with large numbers of marbles for very short periods of time in the approximation task. It is possible that they simply did not have enough time to approximate all of the quantities needed to compute proportion and therefore fell back on a judgmental heuristic in order to make up for the short presentation time. For this reason, my collaborators and I designed a computer-based version of Falk et al. (2012)'s assessment method in order to investigate the influence of procedural and pedagogical factors on children's probabilistic decision making. Importantly, all of the trials included groups of marbles that were easily enumerated by young children and trial images were presented in a self-paced, randomized order.

Findings from Experiments 4.2 and 4.3 suggest that children abandon their probabilistic decision-making strategy when provided with examples that disconfirm their prior beliefs about

probability but not when given a mixed set of confirming and disconfirming examples. Results from the assessment phase indicated that the majority of children in Experiments 4.1 and 4.2 utilized one-dimensional strategies. During the assessment phase, children in the high conflict conditions of both experiments were not significantly different from those in the low conflict conditions. However, during the test phase of Experiment 4.1, children in the high conflict condition (74% correct) performed significantly better than children in the low conflict condition (32% correct) and children in the same condition from Experiment 4.2 were shown to be more likely to change their decision-making strategy than children in the low conflict condition.

5.1.4 Implications for theories of cognitive development

In chapters 2 and 3, I have provided evidence that intuitive representations of probability can be computed rapidly and that the ability to compute probability in this way improves with age. In chapter 4 I have demonstrated that children can reconstruct their understanding of the proportional nature of probability from the proper amount and type of feedback. I see this work as supporting Rational Constructivist theories of cognitive development in two ways. First, Rational Constructivism proposes that humans begin life with a suite of protoconceptual primitives which they can combine and recombine to form ever more complex representations and knowledge (Fedyk & Xu, 2018; Xu, in press). Signature features of the 'number sense' have been observed in human infancy (Feigenson, Dehaene, & Spelke, 2004; Hevia, Izard, Coubart, Spelke, & Streri, 2014; Lipton & Spelke, 2003; Xu & Spelke, 2000; Xu et al., 2005) suggesting that numerical processing is a fundamental component of the human mind. Intuitive probabilistic reasoning appears to be built out of these representations and provides an interesting domain of research on interactions between culture and cognitive development.

Second, Rational Constructivism posits that learners integrate their prior knowledge with new information in a rational way and thus rely on probabilistic cues and statistical inference to do so. Evidence to support this comes from the set of experiments reported in Chapter 4. When children are presented with trials that are randomly sampled from the whole group of 24 trials (low conflict condition), they receive some negative and some positive evidence. The evidence in support of their prior knowledge is mixed and so they stick with their prior intuitions. However, in the 'high conflict' condition, children are provided with negative evidence in conflict with their strategy and positive evidence in favor of their strategy. Since the evidence is no longer mixed, they can more easily override their prior beliefs and come to a more complete understanding of the proportional nature of probability. Interestingly, these results also open up new questions for future research. For example, if infants already have some intuitive understanding of the proportional nature of probability, why is it so hard to teach formal probability in schools? And how does the understanding of mere chance interact with feedback, that is, learners may ignore feedback because they may obtain an unpredicted outcome due to ‘mere chance’ and when do they begin to see that perhaps it is not ‘mere chance’?

5.2 Concluding remarks

It is important to note that this work is the hopeful beginning of a broader research program investigating the role of mental representation in the development of children’s understanding of probability. Although I see this work culminating in an effort to understand the learning processes involved in understanding the proportional nature of probability, I recognize

that further research is needed in order to demonstrate bonafide learning. One pressing issue will be to identify the extent to which children's strategy learning in the random draw task transfers to other forms of non-symbolic and symbolic probability judgments. Does conflicting feedback on a discrete probability judgment task result in strategy change for a continuous probability task? It seems reasonable to assume that the learning we have uncovered in Chapter 4 may have narrow transfer to other forms of discrete and continuous non-symbolic probability judgments. However, it is likely that broader transfer to symbolic probability problems will require direct instruction (Koedinger, Booth, & Klahr, 2013). For this reason, future work will focus on identifying the right combination of discovery learning (i.e. sampling from a distribution) and direct instruction (i.e. formal explanation of proportional relations). This data will hopefully inform future training studies involving both symbolic and non-symbolic probabilistic reasoning.

Bibliography

Acredolo, C., O'Connor, J., Banks, L., & Horobin, K. (1989). Children's ability to make probability estimates: Skills revealed through application of anderson's functional measurement methodology. *Child Development*, 933–945.

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

Allik, J., Tuulmets, T., & Vos, P. G. (1991). Size invariance in visual number discrimination. *Psychological Research*, *53*(4), 290–295.

Alonso, D., & Fernández-Berrocal, P. (2003). Irrational decisions: Attending to numbers rather than ratios. *Personality and Individual Differences*, *35*, 1537–1547.

Alonso-Diaz, S., & Cantlon, J. F. (2018). Confidence judgments during ratio comparisons reveal a Bayesian bias. *Cognition*, *177*, 98-106.

Alonso-Díaz, S., Piantadosi, S. T., Hayden, B. Y., & Cantlon, J. F. (2018). Intrinsic whole number bias in humans. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(9), 1472.

Barth, H., La Mont, K., Lipton, J., & Spelke, E. S. (2005). Abstract number and arithmetic in preschool children. *Proceedings of the National Academy of Sciences*, *102*(39), 14116-14121.

Batanero, C., Chernoff, E. J., Engel, J., Lee, H. S., & Sánchez, E. (2016). Research on teaching and learning probability. In *Research on teaching and learning probability* (pp. 1–33). Springer, Cham.

Bates, Maechler, Bolker, & Walker. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Best Practices, N. G. A. C. for. (2017, december). Common core state standards. Retrieved from http://www.corestandards.org/

Bonato, M., Fabbri, S., Umilta, C., & Zorzi, M. (2007). The mental representation of numerical fractions: Real or integer? *Journal of Experimental Psychology: Human Perception and Performance*, *33*(6), 1410.

Bonny, J. W., & Lourenco, S. F. (2013). The approximate number system and its relation to early math achievement: Evidence from the preschool years. *Journal of Experimental Child Psychology*, *114*(3), 375–388.

Boyer, T. W. (2007). Decision-making processes: Sensitivity to sequentially experienced outcome probabilities. *Journal of Experimental Child Psychology*, *97*(1), 28-43.

Boyer, T. W., Levine, S. C., & Huttenlocher, J. (2008). Development of proportional reasoning: Where young children go wrong. *Developmental Psychology*, *44*(5), 1478-1490.

Boyer, T. W., & Levine, S. C. (2012). Child proportional scaling: Is 1/3= 2/6= 3/9= 4/12? *Journal of Experimental Child Psychology*, *111*(3), 516–533.

Boyer, T. W., & Levine, S. C. (2015). Prompting children to reason proportionally: Processing discrete units as continuous amounts. *Developmental Psychology*, *51*(5), 615–620.

Brainard. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Braithwaite, D. W., & Siegler, R. S. (2017). Developmental changes in the whole number bias. *Developmental Science*.

Bryant, P., & Nunes, T. (2012). *Children's understandling of probability: A literature review (full report)*. Nuffield Foundation.

Castro, C. S. (1998). Teaching probability for conceptual change la enseñanza de la probabilidad por cambio conceptual. *Educational Studies in Mathematics*, *35*(3), 233–254.

Chapman, R. H. (1975). The development of children's understanding of proportions. *Child Development*, *46*, 141–148.

Chiang, W.-C., & Wynn, K. (2000). Infants' tracking of objects and collections. *Cognition*, *77*(3), 169–195.

Dehaene, S. (2011). *The number sense: How the mind creates mathematics.* New York, NY: Oxford University Press.

Dehaene, S., Dehaene-Lambertz, G., & Cohen, L. (1998). Abstract representations of numbers in the animal and human brain. *Trends in Neurosciences*, *21*(8), 355–361.

Denes-Raj, V., & Epstein, S. (1994). Conflict between intuitive and rational processing: When people behave against their better judgment. *Journal of Personality and Social Psychology*, *66*, 819–829.

Denison, S., & Xu, F. (2014). The origins of probabilistic inference in human infants. *Cognition*, *130*(3), 335–347.

De Petrillo, F., & Rosati, A. G. (2019). Rhesus macaques use probabilities to predict future events. *Evolution and Human Behavior*.

DeWind, Adams, Platt, & Brannon. (2015). Modeling the approximate number system to quantify the contribution of visual stimulus features. *Cognition*, *142*, 247–265.

Drucker, C. B., Rossa, M. A., & Brannon, E. M. (2016). Comparison of discrete ratios by rhesus macaques (macaca mulatta). *Animal Cognition*, *19*(1), 75–89.

Durgin, F. H. (1995). Texture density adaptation and the perceived numerosity and distribution of texture. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(1), 149.

Eckert, J., Call, J., Hermes, J., Herrmann, E., & Rakoczy, H. (2018). Intuitive statistical inferences in chimpanzees and humans follow weber's law. *Cognition*, *180*, 99–107.

English, L. D., & Watson, J. M. (2016). Development of probabilistic understanding in fourth grade. *Journal for Research in Mathematics Education*, *47*(1), 28–62.

Fabbri, S., Caviola, S., Tang, J., Zorzi, M., & Butterworth, B. (2012). The role of numerosity in processing nonsymbolic proportions. *The Quarterly Journal of Experimental Psychology*, *65*(12), 2435-2446.

Falk, R., Falk, R., & Levin, I. (1980). A potential for learning probability in young children. *Educational Studies in Mathematics*, *11*(2), 181–204.

Falk, R., Yudilevich-Assouline, P., & Elstein, A. (2012). Children's concept of probability as inferred from their binary choices—revisited. *Educational Studies in Mathematics*, *81*(2), 207–233.

Fazio, L., Bailey, D. H., Thompson, C. A., & Siegler, R. S. (2014). Relations of different types of numerical magnitude representations to each other and to mathematics achievement. *Journal of Experimental Child Psychology*, *123*, 53–72.

Fedyk, M., & Xu, F. (2017). The epistemology of rational constructivism. *Review of Philosophy and Psychology*, 1–20.

Feigenson, Dehaene, & Spelke. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314.

Fischbein, E., & Gazit, A. (1984). Does the teaching of probability improve probabilistic intuitions? *Educational Studies in Mathematics*, *15*(1), 1–24.

Fischbein, E., Pampu, I., & Mânzat, I. (1970). Comparison of ratios and the chance concept in children. *Child Development*, *41*, 377–389.

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education*, 44–63.

Gebuis, T., & Reynvoet, B. (2012). The interplay between nonsymbolic number and its continuous visual properties. *Journal of Experimental Psychology: General*, *141*(4), 642–648.

Ginsburg, N., & Goldstein, S. R. (1987). Measurement of visual cluster. *The American Journal of Psychology*, *100*, 193–203.

Gopnik, A. (2012). Scientific thinking in young children: Theoretical advances, empirical research, and policy implications. *Science*, *337*(6102), 1623–1627.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, *138*(6), 1085.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. Topics in Cognitive Science, 7(2), 217–229.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., … Chan, P. (2016). PsiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, *48*(3), 829–842.

Halberda, & Feigenson, L. (2008). Developmental change in the acuity of the“ number sense”: The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental Psychology*, *44*(5), 1457–1465.

Halberda, Mazzocco, M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, *455*(7213), 665–668.

Halberda, Sires, S. F., & Feigenson, L. (2006). Multiple spatially overlapping sets can be enumerated in parallel. *Psychological Science*, *17*(7), 572–576.

Hevia, M. D. de, Izard, V., Coubart, A., Spelke, E. S., & Streri, A. (2014). Representations of space, time, and number in neonates. *Proceedings of the National Academy of Sciences*, *111*(13), 4809–4813.

Hoemann, H. W., & Ross, B. M. (1971). Children's understanding of probability concepts. *Child Development*, 221–236.

Hurst, M. A., & Cordes, S. (2018a). Attending to relations: Proportional reasoning in 3-to 6-year-old children. *Developmental Psychology*, *54*(3), 428.

Hurst, M. A., & Cordes, S. (2018b). Talking about proportion: Fraction labels impact numerical interference in non-symbolic proportional reasoning. *Developmental Science*, e12790.

Inglis, M., & Gilmore, C. (2013). Sampling from the mental number line: How are approximate number system representations formed?. *Cognition*, *129*(1), 63-69.

Izard, V., Sann, C., Spelke, E. S., & Streri, A. (2009). Newborn infants perceive abstract numbers. *Proceedings of the National Academy of Sciences*, *106*(25), 10382–10385.

Jacob, S. N., & Nieder, A. (2009). Notation-independent representation of fractions in the human parietal cortex. *Journal of Neuroscience*, *29*(14), 4652–4657.

Jacob, S. N., Vallentin, D., & Nieder, A. (2012). Relating magnitudes: The brain's code for proportions. *Trends in Cognitive Sciences*, *16*(2), 157–166.

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

Kahneman, D. (2011). *Thinking fast & slow*. New York, NY: Farrar, Strauss,; Giroux.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. Cognitive Psychology, 3(3), 430–454.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237–251.

Kirkpatrick, L. A., & Epstein, S. (1992). Cognitive-experiential self theory and subjective probability: Further evidence for two conceptual systems. *Journal of Personality and Social Psychology*, *63*, 543–544.

Kleiner, Brainard, & Pelli. (2007). What's new in psychtoolbox-3? *Perception*, *36*.

Koedinger, K. R., Booth, J. L., & Klahr, D. (2013). Instructional complexity and the science to constrain it. *Science*, *342*(6161), 935–937.

Leibovich, T., Katzin, N., Harel, M., & Henik, A. (2017). From "sense of number" to "sense of magnitude": The role of continuous magnitudes in numerical cognition. *Behavioral and Brain Sciences*, *40*.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766.

Libertus, M. E. (2015). The role of intuitive approximation skills for school math abilities. *Mind, Brain, and Education*, *9*(2), 112–120.

Libertus, M. E., Odic, D., & Halberda, J. (2012). Intuitive sense of number correlates with math scores on college-entrance examination. *Acta Psychologica*, *141*(3), 373–379.

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review*, *124*(6), 762.

Lieder, F., & Griffiths, T. L. (2019). Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *The Behavioral and Brain Sciences*, 1-85.

Lipton, J. S., & Spelke, E. (2003). Origins of number sense: Large-number discrimination in human infants. *Psychological Science*, *14*(5), 396–401.

Lourenco, S. F. (2016). How do humans represent numerical and nonnumerical magnitudes? Evidence for an integrated system of magnitude representation across development. In *Continuous Issues in Numerical Cognition* (pp. 375–403). Elsevier.

Matthews, P. G., & Chesney, D. L. (2015). Fractions as percepts? Exploring cross-format distance effects for fractional magnitudes. *Cognitive Psychology*, *78*, 28–56.

Matthews, P. G., & Lewis, M. R. (2017). Fractions we cannot ignore: The nonsymbolic ratio congruity effect. *Cognitive Science*, *41*(6), 1656–1674.

Matthews, P. G., Lewis, M. R., & Hubbard, E. M. (2016). Individual differences in nonsymbolic ratio processing predict symbolic math performance. *Psychological Science*, *27*(2), 191-202.

McCrink, K., & Spelke, E. S. (2010). Core multiplication in childhood. *Cognition*, *116*(2), 204–216.

McCrink, K., & Spelke, E. S. (2016). Non-symbolic division in childhood. *Journal of Experimental Child Psychology*, *142*, 66–82.

McCrink, K., & Wynn, K. (2004). Large-number addition and subtraction by 9-month-old infants. *Psychological Science*, *15*(11), 776–781.

McCrink, K., & Wynn, K. (2007). Ratio abstraction by 6-month-old infants. *Psychological Science*, *18*(8), 740–745.

Mix, K. S., Levine, S. C., & Huttenlocher, J. (1999). Early fraction calculation ability. *Developmental Psychology*, *35*(1), 164-174.

Mix, K. S., Huttenlocher, J., & Levine, S. C. (2002). Multiple cues for quantification in infancy: Is number one of them? *Psychological Bulletin*, *128*(2), 278.

Mix, K. S., Levine, S. C., & Newcombe, N. S. (2016). Development of quantitative thinking across correlated dimensions. In *Continuous Issues in Numerical Cognition* (pp. 1–33). Elsevier.

Möhring, W., Newcombe, N. S., & Frick, A. (2015a). The relation between spatial thinking and proportional reasoning in preschoolers. *Journal of Experimental Child Psychology*, *132*, 213–220.

Möhring, W., Newcombe, N. S., & Frick, A. (2015b). The relation between spatial thinking and proportional reasoning in preschoolers. *Journal of Experimental Child Psychology*, *132*, 213–220.

Moore, C. F., Dixon, J. A., & Haines, B. A. (1991). Components of understanding in proportional reasoning: A fuzzy set representation of developmental progressions. *Child Development*, *62*(3), 441–459.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgments of numerical inequality. *Nature*, *215*(5109), 1519–1520.

NCES (National Center for Education Statistics). (2018, April 27). *Search for public schools: CCD Public school data 2015-2016, 2016-2017 school years*. Retrieved from https://nces.ed.gov/ccd/schoolsearch/index.asp

Ni, Y., & Zhou, Y.-D. (2005). Teaching and learning fraction and rational numbers: The origins and implications of whole number bias. *Educational Psychologist*, *40*(1), 27–52.

Nunes, T., Bryant, P., Evans, D., Gottardis, L., & Terlektsi, M.-E. (2014). The cognitive demands of understanding the sample space. *ZDM*, *46*(3), 437–448.

Odic, D., & Halberda, J. (2015). Eye movements reveal distinct encoding patterns for number and cumulative surface area in random dot arrays. *Journal of Vision*, *15*(15), 1–15.

Odic, D. (2018). Children's intuitive sense of number develops independently of their perception of area, density, length, and time. *Developmental Science*, *21*(2), e12533.

Odic, D., & Starr, A. (2018). An introduction to the approximate number system. *Child Development Perspectives*, *12*(4), 223-229.

Offenbach, S. I., Gruen, G. E., & Caskey, B. J. (1984). Development of proportional response strategies. *Child Development*, 963–972.

O'Grady, S., Starr, A., Griffiths, T. L., & Xu, F. (submitted). Non-symbolic probability judgments rely on integer approximation. *Cognitive Science.*

O'Grady, S., & Xu, F. (2018). Whole number bias in children's probability judgments. In *Proceedings of the 40th annual conference of the cognitive science society*.

O'Grady, S., & Xu, F. (2019). The development of non-symbolic probability judgments in children the development of non-symbolic probability judgements in children. *Child Development*.

O'Grady, S., Griffiths, T. L., & Xu, F. (2016). Do simple probability judgments rely on integer approximation? In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.),

Proceedings of the 38th annual conference of the cognitive science society. Philadelphia, PA: Cognitive Science Society.

Pacini, R., & Epstein, S. (1999). The interaction of three facets of concrete thinking in a game of chance. *Thinking and Reasoning*, *5*(4), 303–325.

Pelli. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

Peng, P., Yang, X., & Meng, X. (2017). The relation between approximate number system and early arithmetic: The mediation role of numerical knowledge. *Journal of Experimental Child Psychology*, *157*, 111–124.

Piaget, J. (1952). *The Child's Conception of Number.* London: Routledge & Kegan Paul Limited.

Piaget, J., & Inhelder, B. (1975). *The origins of the idea of chance in children*. New York, NY: Norton & Company.

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approximate arithmetic in an amazonian indigene group. *Science*, *306*(5695), 499–503.

Rakoczy, H., Clüver, A., Saucke, L., Stoffregen, N., Gräbener, A., Migura, J., & Call, J. (2014). Apes are intuitive statisticians. *Cognition*, *131*(1), 60–68.

R Core Team. (2008). *R: A language and environment for statistical computing. r foundation for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ruggeri, A., Vagharchakian, L., & Xu, F. (2018). Icon arrays help younger children's proportional reasoning. *British Journal of Developmental Psychology*, *36*(2), 313-333.

Saxe, G. B., Gearhart, M., & Seltzer, M. (1999). Relations between classroom practices and student learning in the domain of fractions. *Cognition and Instruction*, *17*(1), 1–24.

Sharma, S. (2015). Teaching probability: A socio-constructivist perspective. *Teaching Statistics*, *37*(3), 78–84.

Siegler, R. S. (2016). Magnitude knowledge: The common core of numerical development. *Developmental Science*, *19*(3), 341–361.

Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology*, *62*(4), 273–296.

Siegler, Strauss, S., & Levin, I. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development*, *46*, Whole No. 189.

Singer-Freeman, K. E., & Goswami, U. (2001). Does half a pizza equal half a box of chocolates?: Proportional matching in an analogy task. *Cognitive Development*, *16*(3), 811–829.

Sophian, C. (2000). Perceptions of proportionality in young children: Matching spatial ratios. *Cognition*, *75*(2), 145–170.

Sophian, C., & Wood, A. (1997). Proportional reasoning in young children: The parts and the whole of it. *Journal of Educational Psychology*, *89*(2), 309-317.

Spinillo, A. G., & Bryant, P. E. (1999). Proportional reasoning in young children: Part-part comparisons about continuous and discontinuous quantity. *Mathematical Cognition*, *5*(2), 181–197.

Starr, A., DeWind, N. K., & Brannon, E. M. (2017). The contributions of numerical acuity and non-numerical stimulus features to the development of the number sense and symbolic math achievement. *Cognition*, *168*, 222–233.

Tecwyn, E. C., Denison, S., Messer, E. J., & Buchsbaum, D. (2017). Intuitive probabilistic inference in capuchin monkeys. *Animal Cognition*, *20*(2), 243–256.

Teglas, E., Girotto, V., Gonzalez, M., & Bonatti, L. L. (2007). Intuitions of probabilities shape expectations about the future at 12 months and beyond. *Proceedings of the National Academy of Sciences*, *104*(48), 19156–19159.

Tourniaire, F., & Pulos, S. (1985). Proportional reasoning: A review of the literature. *Educational Studies in Mathematics*, *16*(2), 181–204.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and 44 probability. Cognitive Psychology, 5(2), 207–232.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. Science, 185(4157), 1124–1131.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315.

U.S. Census Bureau. (2018, April 27). *Quick Facts: Median Household Income (in 2016 Dollars) 2012-2016*. Retrieved from https://www.census.gov/quickfacts/fact/table/US/PST045217

Whalen, J., Gallistel, C. R., & Gelman, R. (1999). Nonverbal counting in humans: The psychophysics of number representation. *Psychological Science*, *10*(2), 130–137.

Wood, J. N., & Spelke, E. S. (2005). Infants' enumeration of actions: Numerical discrimination and its signature limits. *Developmental science*, *8*(2), 173-181.

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

Xu, F., & Kushnir, T. (2012). *Rational constructivism in cognitive development* (Vol. 43). Academic Press.

Xu, F., & Spelke, E. (2000). Large number discrimination in 6-month-old infants. *Cognition*, *74*(1), B1–B11.

Xu, F. (2003). Numerosity discrimination in infants: Evidence for two systems of representations. *Cognition*, *89*(1), B15-B25.

Xu, F., Spelke, E., & Goddard, S. (2005). Number sense in human infants. *Developmental Science*, *8*(1), 88–101.

Xu, F. (in press) Towards a rational constructivist theory of cognitive development. *Psychological Review*.

Yost, P. A., Siegel, A. E., & Andrews, J. M. (1962). Nonverbal probability judgments by young children. *Child Development*, *33*, 769–780.