Do Simple Probability Judgments Rely on Integer Approximation?

Shaun O'Grady (shaun.ogrady@berkeley.edu)
Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Fei Xu (fei_xu@berkeley.edu)
Department of Psychology, University of California,
Berkeley, CA 94720 USA

Abstract

A great deal of research has been conducted on how humans reason about probability, yet it remains unknown what mental computations support this ability. Research on the development of the Approximate Number Sense (ANS) has shown that performance in a magnitude (i.e., estimations of integers) discrimination task is well fit by a psychophysical model (Halberda & Feigenson, 2008). Whether or not estimations of integers plays a role in probability judgments has yet to be investigated. In the present study we use data from two adult experiments as well as results from comparisons of two computational models to investigate the potential relationship between the ANS and probability judgments.

Keywords: Probability; Approximate Number Sense (ANS); estimation

Introduction

Although classic findings in the literature on probabilistic reasoning indicate that adults perform poorly on a wide range of problems involving complex probability (see Kahneman, 2011, for a review), a great deal of evidence suggests that even young children and infants are adept at making correct judgments about simple probability based on proportion (e.g., Denison, Reed, & Xu, 2013; Piaget & Inhelder, 1975). Despite the vast literature on the development of probabilistic reasoning, very little research has been conducted on how these computations are made. In what follows we review findings from research on the development of probabilistic reasoning as well as research on the approximate number sense (ANS) and present findings from two experiments designed to investigate the role of the ANS in adults' judgments about probability based on proportion. We then compare two models proposed to predict probability judgments based on proportion in order to understand the computational process that underlies probability judgments.

Development of Probabilistic Reasoning

Recent evidence suggests that even infants possess powerful statistical reasoning abilities which allow them to predict the outcome of probabilistic events. Researchers have shown that 8-month old infants form expectations about relationships between samples and populations of objects (Xu & Garcia, 2008) and infants as young as 6.5 months expect samples to reflect the ratios of the populations from which those samples were obtained (Denison, Reed & Xu, 2012). These powerful reasoning abilities continue to develop in infancy with older infants making increasingly sophisticated probabilistic judgments about their environment (Denison & Xu, 2009; Gweon, Tenenbaum & Schulz, 2010). As children grow, their judgments about probability become more and more accurate (Piaget & Inhelder, 1975). Recently, Falk, Yudilevich, Assouline & Elstein (2012) found that young children's judgments about simple probability based on proportion are influenced more by the number of favorable events rather than the proportion of favorable to unfavorable events but around 7-8 years of age this changes and children adopt a proportional strategy similar to those used by adults. Based on these findings it seems reasonable to assume that children's improving sense of probability is related to improvements in general number reasoning.

The Approximate Number Sense

Several research teams have shown that both humans and non-human primates have a sophisticated system for making rapid judgments about large and small sets of objects (Whalen, Gallistel & Gelman, 1999; Feigenson, Carey & Hauser 2002, Pica, Lemer, Izard & Dehaene, 2004; Halberda & Feigenson, 2008). This has been termed the Approximate Number Sense (ANS) and has been studied by testing participants' ability to discriminate between two different sets of objects that vary in the magnitude of the difference between the two sets. Several of these studies have found that the acuity of the ANS improves with age and can be characterized by Weber's Law (Halberda & Feigenson, 2008; Pica et al. 2004; Whalen et al. 1999). Accordingly, the ANS is characterized by ratio dependence, meaning that the ability to discriminate between two sets is dependent on the ratio of the magnitudes of those sets.

While there is evidence to suggest that ANS acuity is related to mathematical ability (Halberda, Mazzocco & Feigenson, 2008), some recent evidence has shown that ANS acuity may not influence probabilistic judgments (Patalano, Saltiel, Machlin & Barth, 2015; Winman et al. 2014). Furthermore, some researchers have proposed that rational number reasoning may rely on a set of neural computations that are distinct from that of the ANS (Jacob, Valentin & Neider, 2012). However, it is possible that the computation of rational number comparisons such as proportion and probability, relies on integer estimates. For this reason, it is important to understand the computational process involved in accurate judgments about probability.

Experiment 1: Evaluating Proportions.

Do adults use the approximate number system to reason about the difference between two proportions? To answer this question, we devised a computer-based experiment in which participants are presented with two distributions of red and white marbles and asked to pick the distribution that was most likely to yield a target color marble. If the ANS is recruited when making judgments about probability, then ANS acuity should be negatively correlated with performance on a two-alternative forced-choice task requiring the judgment of probability based on proportion. Additionally, as predicted by Weber's Law, there should be an effect of the distance between the ratios of compared proportions. Finally, if adults are making accurate judgments their responses should be based on the proportion of favorable to non-favorable events rather than the total number of favorable events.

Methods

The methods reported below were originally designed for a study involving young children and for this reason the methods were presented to adults in a child-friendly manner. Only the data from adults is presented below as we are still collecting data from 6- to 12-year-old children.

Participants Forty-eight adult undergraduates (Mean age: 22.62; 37 female) from Psychology classes at UC Berkeley were recruited for participation in this study.

Measures & Stimuli Images were designed using Blender 2.72, 3D animation software (http://www.blender.org/) and consisted of two groups of red and white marbles separated by a black partition. The proportions used are presented in Table 1. All images were presented on a MacBook Pro laptop (OSX; 1280 x 800) using the MatLab programing language with psychToolbox (Brainard, 1997; Pelli, 1997; Kleiner et al, 2007). In order to measure the acuity of participants' approximate number sense, participants played Panamath a game designed to measure the acuity of the ANS (http://panamath.org/) for 10 minutes after the probability task.

(a) (b)

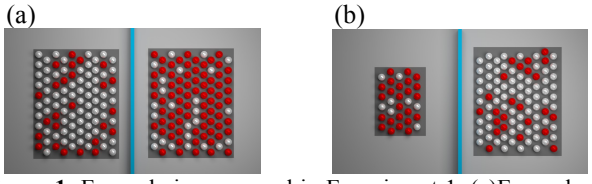

Figure 1: Example images used in Experiment 1. (a)Example of a total equal trial. (b) Example of a target equal trial

Procedure Participants sat about 60 centimeters from the laptop on which the probability game and Panamath task were presented. On the initial screen, participants saw a Big Bird character with two bags on either side and were told that Big Bird really likes red marbles. They were then told that Big Bird is going to close his eyes and take a marble from one of two groups of marbles. Their job was to tell Big Bird which group to take a marble from to help him collect as many of his favorite color marbles as possible. Participants then played four practice trials in which they saw a group of red marbles on one side and a group of white marbles on the other side.

Two types of trials were used for each ratio of ratios listed in Table 1. On 'total equal' trials both groups of marbles contained 100 marbles while on 'target equal' trials the number of Big Bird's favorite color marbles in each group was equal. See Figure 1 for an example image of the two trial types. If participants are using a strategy based on choosing the distribution with the greatest amount of favorable marbles, they should perform at around chance levels on target equal trials.

All of the images were presented on screen for 500ms in order to prevent participants from counting the marbles. Participants played the game for 40 trials presented in one of two orders for both red and white target colors. Trial order and target color were counterbalanced across subjects. Adults participants were told that the game was originally made for young children and therefore used a lot of child friendly language.

Table 1: Ratios of marbles used in Experiment 1.

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

Note: 'Bag 1 prop.' etc, indicates the proportion of favorable marbles.

Results

Analyses of general performance revealed the average percentage correct to be 94.9% (*SD* = 0.04%) and this was significantly greater than that expected by chance ($t = 64.93$, $df = 47$, $p < .001$). The average Weber Fraction (WF) for the 30[1] adults tested on Panamath was 0.162 (*SD* = 0.0058). Importantly, WF was not statistically significantly correlated with general performance in the probability game (Pearson's $r = -0.32$, $t = -1.8036$, $df = 28$, $p = .082$).

Generalized Linear Models with Mixed effects (GLMMs) were used to predict performance based on age, ratio of ratios, and trial type with participant entered as a random effect and found no main effect or interaction effects for gender, trial order or target color. The model predicting performance based on the ratio of ratios and trial type with an interaction ($AIC_{RR*trial} = 772.04$) outperformed the null model ($AIC_{null} = 837.15$) ($\chi^2 = 71.105$, $df = 3$, $p < .001$), the

[1] The scores of 18 participants were excluded from these analyses because they played the Panamath game for fewer than 8 minutes which is the required amount of time to acquire an accurate measure of number sense acuity.

model predicting performance based on ratio of ratios (AIC_{RR} = 822.81) (χ^2 = 54.7, *df* = 2, *p* < .001), age (AIC_{age} = 839.15) (χ^2 = 71.105, *df* = 2, *p* < .001), trial type (AIC_{trial} = 824.72) (χ^2 = 56.674, *df* = 2, *p* < .001), ratio of ratios and trial type ($AIC_{RR+trial}$ = 810.22) (χ^2 = 40.177, *df* = 1, *p* < .001), ratio of ratios, trial type, and age ($AIC_{RR+trial+age}$ = 812.22) (χ^2 = 40.176, df = 0, *p* < .001). Average performance by trial type and ratio of ratios is plotted in Figure 2.

Analysis of the coefficients revealed that as the ratio of ratios increased the chances of a correct response increased by 4%. The chance of a correct response on total equal trials was 88% less likely than for target equal trials and the effect of ratio of ratios was 4.9 times greater for total equal trials than for target equal trials.

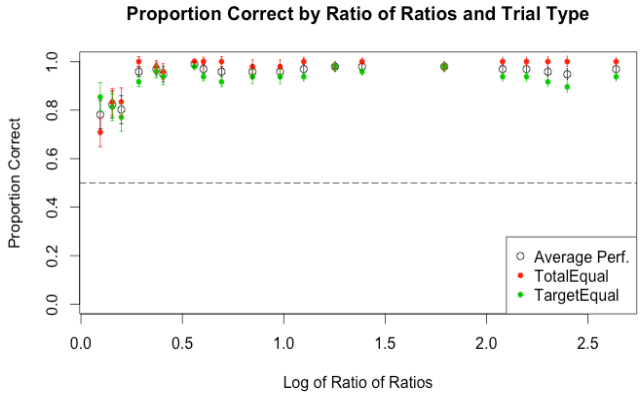

Figure 2: Percentage correct by trial type and ratio of ratios.

Discussion

These results suggest that the ANS may not be used in judgments of probability based on proportion: there was not a statistically significant correlation between participants' Weber fraction and general performance in the probability discrimination task, with a sample size (*n* = 30) where a large effect (*r* > 0.5) would have been detected (Cohen, 1988). Interestingly, there was an effect of ratio of ratios on performance in this task and this effect was in the expected direction: as the ratio of ratios increases so did performance, a finding that mirrors the distance effect reported in the ANS literature.

The GLMMs also revealed that performance improved on trials in which the number or target marbles is equal suggesting that participants' judgments were influenced by the number of favorable marbles. However, it should be noted that performance on both target equal and total equal trials was significantly above chance indicating that participants did not adopt a strategy based solely on the number of favorable marbles but rather that performance was enhanced on trials in which the target number of marbles was equal indicating that some participants may have adopted a strategy of avoiding the choice with the most red on some trials.

Several questions arise in response to these findings. First, and most importantly, it is interesting that there was no significant correlation with ANS acuity but it remains unknown if this is true of children as well. It is possible that adults have learned a non-numerical strategy and therefore no longer use their ANS when making probabilistic judgments. Future research will investigate the influence of ANS acuity on children's probability judgments. Second, the astute reader will notice that the ratio of ratios that were used in Experiment 1 were mostly comparisons of favorable (above 50% chance) distributions to unfavorable (below 50% chance) distributions. It seems likely that different results may be found when comparing favorable distributions to other favorable distributions as well as comparing unfavorable to other unfavorable distributions.

Another issue with the design of Experiment 1 arises from the analysis of trial type in which there was enhanced performance for target equal trials. Using the current design it is impossible to tell if participants were relying solely on approximations of either target or non-target marbles or whether they were reasoning about the proportion of target to non-target marbles. Trials in which the distribution with a lower proportion actually has a higher number of favorable marbles would show such an effect. Finally, it is impossible to tell to what extent participants' judgments are influenced by the proportion of area of the two colors rather than the proportion of the number of marbles, given that all the marbles were the same size. Considering the lack of correlation with ANS acuity in the adult sample, it is possible that participants are making proportional judgments based on area rather than numerosity. Each of these points is addressed by design modifications made in Experiment 2.

Experiment 2: Area, number & chance.

Several changes were made to the design of Experiment 2 in order to address the potential limitations of Experiment 1 and explore the possibility that participants are not recruiting the ANS when making judgments but are instead using other possible alternative strategies. The results we present here constitute pilot data for a more comprehensive study. The full version of the study will include more trials with a wider range of ratio comparisons.

Methods

Participants 19 undergraduates (Mean age: 21 years (SD = 1.94 years); 14 female) enrolled in psychology classes at U.C. Berkeley participated in this experiment.

Measures & Stimuli New ratio of ratios comparisons were chosen to include favorable vs unfavorable (F:U), favorable vs favorable (F:F), favorable vs 50% chance (F:C), 50% chance vs unfavorable (C:U), and unfavorable vs unfavorable (U:U) comparisons similar to Drucker, Rossa and Brannon (2016).

Three trial types were used for each of the ratio of ratios. On 'total equal' trials the total number of marbles in both distributions was equal while on 'number vs proportion' trials there were more favorable marbles in the distribution with the lower proportion. As in Halberda and Feigenson

(2008), 'area anti-correlated' trials included larger favorable marbles in the lower proportion distribution and smaller favorable marbles in the higher proportion distribution such that the proportion of area of favorable color in the lower proportion distribution was equal to the proportion of the number of marbles in the higher proportion distribution and vice versa. See Figure 3 for example images of trial types.

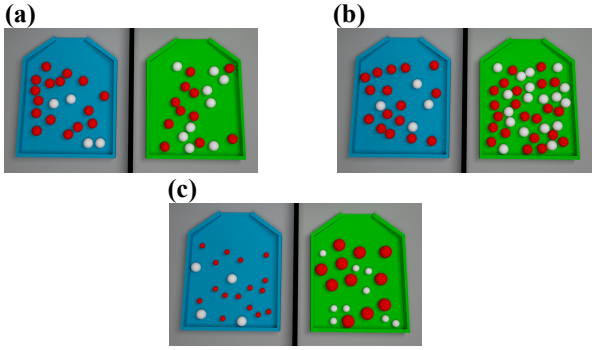

Figure 3: Example images used in Experiment 2. (a) Example of a total equal trial. (b) Example of a number vs proportion trial. (c) Example of an area-anticorrelated trial.

Procedure Images were presented using the same procedure as in Experiment 1 and the instructions were exactly the same except that participants were no longer asked to help Big Bird but instead were asked to choose the box most likely to yield a red marble. During practice trials participants were shown that the trays containing the marbles were dumped into two different boxes which were then shaken up. This was done in order to reduce the use of spatial cues that could influence the participant's selection. box they chose would tip over and dispense a single marble. Participants were also told that the size of the marble did not matter and the large marbles are just as likely to fall out of the box as small marbles. Finally, participants played the Panamath game after the probability task for 10 minutes in order to measure their ANS acuity.

Results

Participants' average performance is presented in figure 4. Mean overall percentage correct was 70.3% (*SD* = 14.9%), and this was significantly better than chance (t = 5.936, *df* = 18, $p < .001$)). The average Weber Fraction (WF) as measured by Panamath was 0.156 (*SD* = 0.029) and participant's WF was not statistically significantly correlated with general performance in the probability game (Pearson's $r = -0.39$, $t = -1.762$, $df = 17$, $p = .096$).

As with Experiment 1, GLMMs were used for the analyses. Model comparisons revealed that the model with the best fit was that which predicted performance based on ratio of ratios, trial type, and the interactions between ratio of ratios and trial type ($AIC_{RR*trial} = 576.33$). This model outperformed the null model ($AIC_{null} = 677.35$) ($\chi^2 = 111.02$, $df = 5$, $p < .001$), as well as models predicting performance based on the ratio of ratios ($AIC_{RR} = 652.97$) ($\chi^2 = 84.64$, $df = 4$, $p < .001$), trial type ($AIC_{trial} = 608.78$) ($\chi^2 = 38.45$, $df = 3$, $p < .001$), and ratio of ratios and trial type ($AIC_{RR+trial} = 580.4$) ($\chi^2 = 8.07$, $df = 2$, $p = .018$). Gender, age, comparison category, trial order and target color were not significant predictors of performance.

Analyses of the coefficients of the superior model revealed that as the ratio of ratios increased participants were more than twice as likely to make a correct choice. Both the number vs proportion and the area anti-correlated trials had a negative effect on the likelihood of a correct response with number vs proportion trials decreasing the chances of a correct response by 93.2% and area anti-correlated trials by 12.7%. Interestingly, the interaction terms indicated that the effect of ratio of ratios increased by 4% for number vs proportion trials and decreased by 52.4% for area anti-correlated trials.

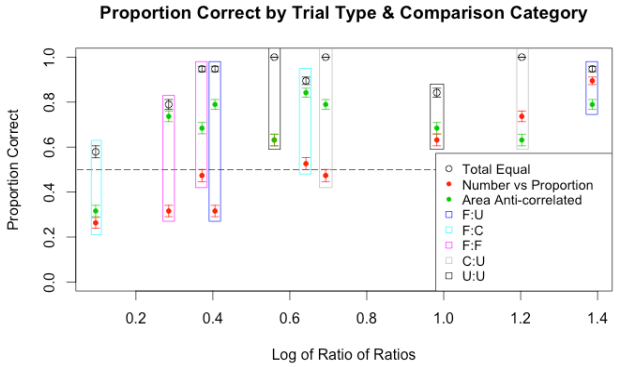

Figure 4: Proportion correct by trial type, ratio of ratios, and comparison category

Discussion

Not surprisingly, performance on this task was much worse than performance in Experiment 1 and this is probably due to the fact that we included many more difficult trials. In particular, the number vs. proportion and area anti-correlated trials were meant to be more difficult than the target equal trials used in Experiment 1. The interaction between ratio of ratios and trial types suggest that participants are using an area calculation as part of their estimations and this is interesting given the lack of an effect of area in the ANS literature (Halberda & Feigenson, 2008). Importantly, the comparison category did not seem to affect participant's judgments which may indicate that adult participants are able to compute an accurate probability estimate regardless of the likelihood of obtaining a target event. However, since there are only two of each comparison category and the comparisons are not evenly spaced throughout the ratios of ratios used in the study, the effect of qualitative distinctions between comparisons must remain open to speculation.

Interestingly, performance was still above chance even with the more difficult trials used in Experiment 2 and there was still no correlation with ANS acuity even though there was still a significant effect of the ratio of ratios. It is still too early to definitively rule out the role of the ANS in probability judgments since participants may be using some computation applied to their estimates of the number of

target and non-target marbles. For example, the participant might estimate the total number of target marbles as well as the total number of marbles for both distributions. Since ANS estimates are inexact, even for those who have with high number sense acuity, an error in any one of these estimates could lead to an incorrect judgment. For this reason it is necessary to understand the computations that lead to accurate performance as well as the strategies that people may adopt when reasoning about the probability of discrete binary outcomes.

Models of Probability Discrimination

Several studies investigating the approximate number sense (Halberda & Feigenson, 2008; Pica et al. 2004) have used computational models based on signal detection theory in the psychophysical literature (Green & Swets, 1966). The psychophysics model assumes that participants' approximations form a Gaussian distribution on a mental number line around a mean close to the actual number of objects being estimated and a standard deviation given by the participants' number sense acuity, or Weber fraction, multiplied by the mean. When presented with two groups of objects and asked which has more the participants' estimates form two Gaussian distributions which overlap based on the distance between their means and the size of their standard deviations. The area of overlap can be thought of as the probability of a subject making an error, which can be modeled using the psychophysical function based on the CDF of the Gaussian distribution:

$$P(R_c) = \Phi\left(\frac{\mu_1-\mu_2}{\sqrt{2w}\sqrt{\mu_1^2+\mu_2^2}}\right) = \frac{1}{2}erfc\left(\frac{\mu_1-\mu_2}{\sqrt{2w}\sqrt{\mu_1^2+\mu_2^2}}\right) \quad (1)$$

where $P(R_c)$ is the probability of a correct response, w is the subjects' Weber fraction, $\mu_1$and μ_2 are the means of the two numbers being estimated, and Φ is the Cumulative Distribution Function (CDF) of the Gaussian distribution.

This model assumes that participants estimate probability as an exact numerical value between 0 and 1. Although Equation 1 seems to be a plausible model for how people are making judgments about probability based on comparisons of two binary distributions, assessment of the likelihood of this model reveals that it is not well fit to the data. Figure 5 provides the model predictions alongside the data from Experiments 1 and 2.

The probability of selecting a target color marble from the distributions used in Experiments 1 and 2 can be represented as rational numbers between 0 and 1. If the subjects in our study are using their ANS to calculate probabilities they would need to approximate the number of favorable objects and divide this by the total number of all objects in the distribution which could be modeled using the distribution of a ratio of two random variables. Although a general form for calculating the probability density function for this distribution exists (Hinkley, 1969), a model using assumption distribution makes similar predictions as the model assuming a Gaussian distribution with the ratio comparisons that are reported for Experiments 1 and 2 above.

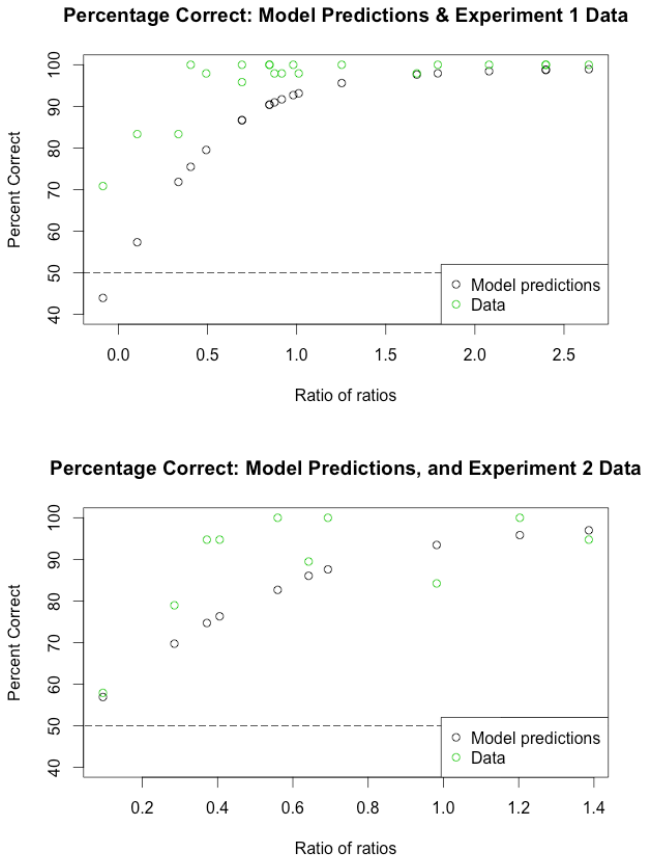

Figure 5: Model predictions for the Gaussian model along with data from Experiments 1 and 2. (a) Predictions and data for Experiment 1. (b) Predictions and data for Experiment 2.

General Discussion

Using a combination of experimental evidence and computational modeling we sought to demonstrate the role of the approximate number sense in adult participants' judgments of probability based on proportion. In two experiments we demonstrated that adults' ANS acuity was not statistically significantly correlated with performance on a probability discrimination task. Since a large correlation would have been detected in these experiments, these results suggest that the ANS is not recruited when people make judgments about probability based on proportion.

It is possible that participants' judgments about probability comparisons are more heavily influenced by qualitative factors such as trial type and comparison category. Future work will include larger numbers of more varied comparisons in order to more thoroughly investigate the factors influencing probability judgments. In addition, we plan to use more complex models that can account for

both quantitative and qualitative factors in judgments of probability based on proportion.

Acknowledgments

We would like to thank Falk Lieder and Dylan Daniels of U.C. Berkeley for their suggestions on this manuscript. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. (DGE 1106400)

References

Brainard, D. H. (1997) The Psychophysics Toolbox, *Spatial Vision 10*:433-436.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Denison, S. & Xu, F. (2010) Twelve- to 14-month-old infants can predict single-event probability with large set sizes.*Developmental Science, 13,* 798-803.

Denison, S., Reed, C., & Xu, F. (2013). The emergence of probabilistic reasoning in very young infants: evidence from 4.5-and 6-month-olds. *Developmental Psychology*, *49*(2), 243.

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.

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.

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, *107*(20), 9066-9071.

Halberda, J., & 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.

Halberda, J., Mazzocco, M., & Feigenson, L. (2008). Individual differ- ences in non-verbal number acuity correlate with mathematics achievement. *Nature*, 455, 665–668.

Hinkley, D. V. (1969). On the ratio of two correlated normal random variables. *Biometrika*, *56*(3), 635-639.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kleiner M, Brainard D, Pelli D, 2007, "What's new in Psychtoolbox-3?" Perception 36 ECVP Abstract Supplement.

Patalano, A. L., Saltiel, J. R., Machlin, L., & Barth, H. (2015). The role of numeracy and approximate number system acuity in predicting value and probability distortion. *Psychonomic Bulletin & Review*, *22*(6), 1820-1829.

Pelli, D. G. (1997) The VideoToolbox software for visual psychophysics: Transforming numbers into movies, *Spatial Vision 10*:437-442.

Piaget, J., & Inhelder, B. (1975). *The origin of the idea of chance in children. (Trans L. Leake, P. Burrell & HD Fishbein)*. WW Norton.

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

Winman, A., Juslin, P., Lindskog, M., Nilsson, H., & Kerimi, N. (2014). The role of ANS acuity and numeracy for the calibration and the coherence of subjective probability judgments. *Frontiers in Psychology*, *5*.

Xu, F., & Denison, S. (2009). Statistical inference and sensitivity to sampling in 11-month-old Infants. *Cognition*, *112*(1), 97-104.

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