Adults' guesses on probabilistic tasks reveal incremental representativeness biases

Habiba Azab[1] **(hazab@mail.bcs.rochester.edu)**
David Ruskin[1] **(druskin@mail.bcs.rochester.edu)**
Celeste Kidd[1,2] **(celestekidd@gmail.com)**
[1]Department of Brain and Cognitive Sciences and [2]Center for Visual Sciences
University of Rochester, Meliora Hall, Rochester, NY, 14627

Abstract

Participants in most binary-choice tasks with multiple trials tend to probability-match (Vulkan, 2000) — i.e., provide responses that match the probability distribution of the presented population. Given a single trial, however, participants usually choose the majority option (James & Koehler, 2011). By using a method that visually presents the probabilities of the two competing options, we examine responses when participants are given only a single trial, and initial responses when participants are given multiple trials. While we still observe aggregate probability-matching in the multiple-trial condition, we find robust sequence effects in participants' initial responses, including robust maximizing behavior on the first response. This suggests that both maximizing in single-trial experiments and aggregate probability-matching in multiple-trial ones can be explained by a single, underlying mechanism; one that seeks to provide a representative sample at each point during sequence generation.

Keywords: Decision making; statistics; psychology; human experimentation; probability-matching; maximizing.

Participants Adopt Sub-Optimal Strategy in Simple Binary-Choice Tasks

Economics, as well as daily dealings and discourse, rest upon a fundamental assumption of human rationality: that individuals will make choices that maximize their expected benefit. However, in one of the simplest experimental setups—the binary-choice task—adult participants fail to conform to this assumption. In this task, participants are repeatedly asked to choose between two options, each exclusively offering a reward with a certain probability (for example, an orange button is rewarding on 70% of trials and a green button on the remaining 30%). These probabilities are fixed, and the distribution of rewards across trials is independent of reward history or participants' performance. The optimal strategy—the one that maximizes total payout—is to select the option with the highest probability (the majority or dominant response) on every trial (e.g. given 10 trials, press the orange button exclusively). Yet participants tend to probability-match, matching the distribution of their responses to the reward distribution of the options (e.g. pressing the orange button 7 times and the green button 3 times). This result is well established (Vulkan, 2000 for a comprehensive review; Grant, Hake & Hornseth, 1951; Hake & Hyman, 1953; Gardner, 1957; Rubinstein, 1959; West & Stanovich, 2003). Previous research has demonstrated that adults may be coaxed more towards maximizing by using higher payoffs (Toda, 2009), risk of losing gains (Siegel & Goldstein, 1959; Siegel, 1961), detailed feedback (Friedman & Massaro, 1998; Birnbaum & Wakcher, 2002), an extensive number of trials (Edwards, 1965), emphasizing the random nature of the task (Goodnow, 1955), or some combination of these (Shanks, Tunney & McCarthy, 2002). Even then, however, not all adults maximize.

Shanks, Tunney and McCarthy (2002) argue that the typical experimental setup, with a large number of rapid, successive binary choices—typically 100 to 400 (Vulkan, 2000) and as many as 1,000 (Edwards, 1961)—might be ecologically questionable. Humans are not typically faced with the same choice hundreds of times on a single day, especially with static reward probabilities. They predict that participants might maximize if given fewer, infrequent choices. To test this, James and Koehler (2011) presented participants with a sequence of ten statistically identical choices in the guise of ten different game setups—emphasizing the unique nature of each gamble. They predicted that participants would be more likely to maximize in the single-shot games. True to this intuition, participants were more likely to maximize in the unique-gambles condition and probability-match in the repeated-gambles condition.

James and Koehler (2011) suggest that the mechanism underlying adults' guessing behavior depends on their expectations about how many guessing opportunities they will have. In essence, adults could employ two distinct strategies depending on how many guesses they will have to make—maximizing when they have only one, and probability-matching when they have many. We propose an alternative hypothesis: the *incremental-representativeness hypothesis*, inspired by Kahneman and Tversky's representativeness bias (1972). This hypothesis proposes a trial-by-trial mechanism where participants provide a sample that is *incrementally* representative of the population—i.e. at each point of the experiment. We believe this hypothesis is capable of explaining maximizing behavior given a single trial, probability-matching behavior over several trials, and order effects observed in the latter setting, particularly over participants' first few responses.

The Incremental-Representativeness Hypothesis

More specifically, the incremental-representativeness hypothesis suggests that participants observe the distribution of the population and, at each guess, aim to make their previous responses representative of that population. This hypothesis makes particularly strong predictions for responses in the first few trials of an experiment. At the first guess in a series, participants are expected to maximize—i.e. choose the more probable option—regardless of how many guesses are to come afterwards. This is because, given only one trial, the majority choice is more representative of a skewed distribution than the minority one. Over several participants, order effects would be most strongly observed during initial trials—more specifically, the minimum number of samples necessary to fully describe the distribution (e.g. 3 in a 2:1 distribution, 5 in a 2:3 distribution, 100 in a 99:1 distribution, etc).

In contrast, trial-by-trial probability-matching predicts that the likelihood of a particular response remains constant throughout the experiment, and matches the probability of that response in the population. For example, in a population of 70% orange and 30% green gumballs, the trial-by-trial probability-matching hypothesis predicts that participants are 70% likely to guess orange and 30% likely to guess green at each trial during the experiment, regardless of the order of that trial in the overall sequence of responses. The underlying mechanism behind this hypothesis is best visualized by picturing the participant flipping a weighted-coin: the coin is always (say) 70% likely to turn up heads and 30% likely to turn up tails, regardless of how many times it is flipped, or which flip in the sequence we are at.

Note that both the incremental-representativeness hypothesis and the trial-by-trial probability-matching hypothesis would give rise to the same aggregate sample statistics, where the proportions of the two available alternatives match their proportions in the given population. These hypotheses differ, however, in the predictions they make at the level of a single-trial. While the incremental-representativeness hypothesis can explain both aggregate probability-matching (given several trials) and maximizing (given only one trial), probability-matching at the single-trial level fails to explain maximizing given only one trial, and predicts no significant order effects.

We tested our hypothesis with a simple binary-choice task (Experiment 1). We randomly assigned adult participants to either a single-trial or a multiple-trial condition. We predict that adults will maximize on the first trial, regardless of whether the first guess was their only guess. We also predict that participants will probability-match over the entire sequence of given trials, although not on each individual trial, in the 10-trial condition. Alternatively, participants could probability-match on each individual trial of the experiment (including the first), supporting the trial-by-trial probability-matching hypothesis, and suggesting that participants utilize different guessing strategies depending on the number of trials they expect to be given.

Participants in our single-trial condition, as a group, maximized—replicating James and Koehler's (2011) findings and providing an experimental estimate of maximizing behavior. In line with our predictions, participants in the multiple-trial condition also maximized on the first trial, rather than probability-matched. This result suggests that maximizing behavior is not elicited by unique gambles, but by the first gamble in general. Responses in the first three trials of the experiment suggest that, rather than probability-matching on individual trials, participants seemed to provide incremental samples that were representative of the population not only after all trials were completed, but after each successive trial.

We further tested this hypothesis by asking an independent pool of participants to evaluate the likelihood of different sequences being drawn at random from the same population (Experiment 2). We find that the sequences participants tended to provide in Experiment 1 were rated highly by participants in Experiment 2. These results further support the incremental-representativeness hypothesis.

Experiments

We chose to present our binary-choice population using a static image of a gumball machine, filled with 70% orange gumballs and 30% green gumballs. We chose this particular task design for three main reasons. First, unlike the lights task (Anderson, 1960; Derks & Paclisanu, 1967)—in which the probability distribution is only apparent after some number of trials,—the gumball machine provides an immediate, non-sequence-dependent display of the distribution of the outcomes. This offers three distinct benefits: 1) it places fewer demands on working memory, 2) it avoids the risk of introducing unintentional and entirely random sequential patterns in the initial presentation of the population's distribution, and 3) unlike most previous experimental setups, it requires no training to learn the distribution of the population, allowing us to look at participants' responses from the very first trial onwards, where we are most likely to see the order effects predicted by our hypothesis[1].

Second, the gumball machine makes the primary assumptions of the task intuitive and reasonable: that outcomes are random and independent, and not subject to human control. Similar experimental setups have been used in recent studies with children, in which probabilities were represented using ping pong balls in clear containers (Xu & Garcia, 2008; Xu & Denison, 2009), and more recently using a gumball machine (Denison & Xu, 2014). Finally, we chose to provide no feedback until the end of the experiment to eliminate both pattern-seeking and 'win-stay, lose-shift' behaviors.

However, this framework is not without certain drawbacks. Drawing gumballs from a gumball machine is a process of sampling without replacement. We thus did not give participants any more than 10 trials, ensuring that sampling from the gumballs would not skew the population's distribution significantly. Also, like many modern experiments, ours relies on a computer-simulated task, not an actual gumball machine. This could interfere with participants' perceptions of randomness. Ideally, this experiment would be replicated with an actual gumball machine in a natural setting with replacement, to dispel any notions the participants might have about our interfering with the outcome of trials.

It is possible that the order effects we observe in these

[1] While past experiments have reported participants' responses in the initial trials, these are essentially random as participants are in the process of learning the underlying population distribution.

data are somehow due to the specifics of the task—such as the sequential nature of responding, by which participants give their guesses one at a time and are not shown what they guessed previously, nor are they given any feedback (to match the 1-trial and 10-trial conditions of this experiment as closely as possible). To eliminate this hypothesis, we ran a follow-up experiment. In Experiment 2, we ask a different group of participants to evaluate the likelihood of various sequences of gumballs, while emphasizing the order in which they come out of the machine through a simple visual animation. This relieved participants of the need to remember previous guesses, and allowed us to determine whether participants rate highly the same sequences that they are likely to respond with in the first task. This manipulation also gives us an experimental measure of what sequences participants actually consider to be representative.

Method

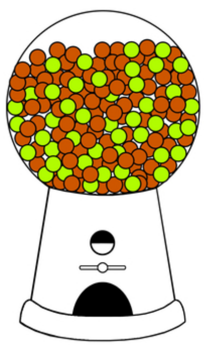

Figure 1: Gumball machine image used in both experiments

Experiment 1

173 participants[2] (mean age = 32 years, range = 18-66 years old), recruited through Amazon Mechanical Turk, were presented with the image of a gumball machine (Figure 1), in which 70% of gumballs are orange and 30% are green[3]. They were asked to look at the machine carefully and guess what color gumball would come out of it next. They were also told that they would be given a bonus based on their good performance and feedback. Participants were assigned to conditions at random; 86 participants were allowed only 1 guess, while 87 participants were allowed 10 guesses. The number of guesses remaining was shown throughout the experiment. Two gumball-shaped (i.e., round) buttons—one orange and one green—were displayed at the bottom of the screen; participants were required to choose one, then click the "Guess" button to make their guess and, in the 10-trial condition, proceed to the next trial. The locations of the buttons indicating the two choices were counterbalanced; in the 10-trial group, the locations were also switched at random between trials. At the end of the experiment, participants were given "feedback" about their performance; this was, in fact, unrelated to the guesses they made and standardized for all participants ("You guessed correctly!" for 1-trial participants, and "You guessed correctly on 8 trials!" for 10-trial participants) [4]. Participants were then asked to estimate the percentage of orange and green gumballs in the machine, and to provide a brief description of the strategy they used to make their choices.

Experiment 2

In a follow-up experiment, 154 participants[5] (mean age = 34 years, range = 19-68 years old), also recruited through Amazon Mechanical Turk, were presented with the same gumball-machine image shown in Figure 1. They were then shown a sequence of 1 to 4 gumballs, asked to look at the machine carefully, and then rate the likelihood of the shown sample on a 7-point scale (1: least likely, 7: most likely). They were presented with all 30 possible color permutations of 1, 2, 3 or 4 gumballs, as well as 4 catch trials with purple (distractor) gumballs to assess attentiveness. (Since the image contained no visible purple gumballs, attentive participants would be expected to assign these sequences low probabilities.) After all trials were completed, participants were asked to estimate the percentage of each color in the machine.

Results

Experiment 1

Participants estimated the percentage of orange gumballs at 66.07% [6] (median = 65%, standard deviation: 7.58%). This was slightly lower than the actual percentage of orange gumballs (70%). This could be due to differences in the brightness of the colors presented, or indicating that participants might be estimating the closest ratio (2:1), rather than a percentage. For all subsequent analyses, we will use the mean proportion of orange gumballs that participants predicted (0.66).

[2] 33 additional participants were excluded from all analyses due to failure to understand the task or inattentiveness: either by estimating both options to be equally distributed, or by estimating the percentage of green gumballs to be higher than that of orange ones.

[3] We never explicitly refer to the colors of the gumballs as "orange" or "green", since different monitors might display these shades differently. For all responses, we ask participants to click on a button of the same color as the gumball.

[4] Feedback at this stage of the experiment could not affect participants' performance, since they had already completed all trials.

[5] 57 additional participants were excluded from all analyses due to failure to understand the task or inattentiveness. Given that this was a longer task, we define more stringent criteria for subject inclusion. 13 participants were excluded for rating combinations with a purple (distractor) gumball at 4 or higher (where 7 = 'very likely'), 6 more participants were excluded for providing estimates of the percentage of orange and green gumballs that did not add up to 100 +/- 5%, an additional 3 participants were excluded for estimating the percentage of green gumballs to be higher than that of orange gumballs, and the rest (35 participants) were excluded for estimating the percentage of green gumballs to be equal to that of orange gumballs. While we realize this is a high percentage of participants to exclude, all reported effects remain even when no participants are excluded.

[6] 15 participants provided estimates of orange and green gumballs that did not add up to 100%, but seemed to indicate a ratio rather than a percentage. These responses were converted to percentages.

We note, however, that the pattern of results remains the same when the actual proportion (0.7) is used instead.

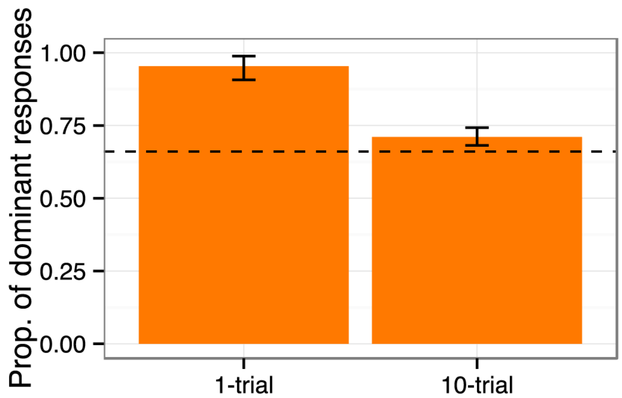

Figure 2: Comparing the mean of majority (orange) responses in the 1-trial and 10-trial conditions. Dashed line indicates participants' estimate of the proportion of orange gumballs (66%). Error bars are bootstrapped 95% confidence intervals.

Results show that, overall, participants maximized in the 1-trial condition, and probability-matched in the 10-trial condition. Figure 2 shows the mean proportion of majority responses from participants in the 1-trial and 10-trial conditions; these means were significantly different (Wilcoxon rank-sum test: W=6601, $p < 10^{-15}$). The mean proportion of majority responses in the 1-trial condition was significantly larger than 0.66 (Wilcoxon signed rank test: V=3403, $p < 10^{-12}$), while the same proportion from the 10-trial condition was not significantly different from 0.66 (Wilcoxon signed rank test: V=2176, $p = 0.265$). In the 10-trial condition, we were also able to compare the proportion of majority responses per subject against their own estimated fraction of orange gumballs. These do not differ significantly, either (Wilcoxon signed rank test: V=1279, $p = 0.110$). This indicates that while, overall, participants in the 10-trial condition are probability-matching, participants in the 1-trial condition are maximizing. This replicates previous findings (James & Koehler, 2011).

Figure 3 shows the proportion of participants who gave each possible number of majority responses. This plot shows that probability-matching is not a behavior observable only at the population level; the majority of our participants probability-match. It is worth noting, however, that a number of participants (13) maximized—choosing the majority response exclusively. We re-ran all analyses while excluding this population of maximizers, and all results remain qualitatively the same.

We then examined the proportion of majority responses in each individual trial of the 10-trial condition (Figure 4). If participants were strictly probability-matching at the trial level, we would expect to see all of these proportions at or around 0.66. However, the fraction of majority responses in the first and second trials were significantly higher than 0.66 (Wilcoxon signed rank test, 1st trial: V=3240, $p < 10^{-9}$; 2nd trial: V=2850, $p < 10^{-4}$). This fraction then decreased below 0.66 on the third trial (Wilcoxon signed rank test, V=1275, $p = 0.0052$), after which sequence effects seem to fade. Note that 3 trials are sufficient to represent a roughly 2:1 distribution.

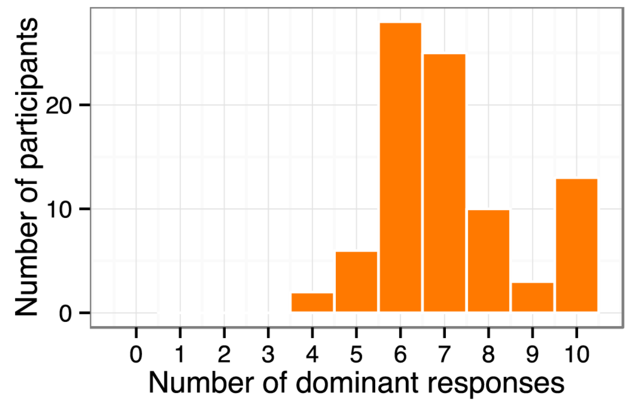

Figure 3: Proportion of 10-trial participants who gave 'x' majority-option (orange) responses over entire experiment (n=87)

Finally, we used our estimate of maximizing behavior from the 1-trial condition to analyze behavior on the first trial in the 10-trial condition. We found no significant difference between the proportion of majority responses chosen on this trial, and on the only trial in the 1-trial condition (Figure 5; Wilcoxon rank-sum test: W=3868, $p = 0.364$). This suggests that participants utilize the same strategy on the first trial of an experiment, regardless of whether they have more trials left.

Experiment 2

The survey data show a general correspondence between the initial sequences (length 1-4) participants are most likely to respond with in Experiment 1, and ones they tend to rate as most likely in Experiment 2. Figure 6 shows a plot of the average normalized rating per sequence[7], and the likelihood of that sequence in the dataset previously discussed. Note particularly the high average ratings for '1'('orange'), '11'('orange, orange'), and '110'('orange, orange, green'). Compare these to the relatively low rating for '011'('green, orange, orange'), which is objectively as likely as '110', but rated much lower. This suggests a strong order bias in both participants' responses and evaluations.

Discussion

We presented two groups of participants with a binary-choice task in which they had to predict which color gumball would

[7]Results are qualitatively similar when raw, un-centered ratings are used.

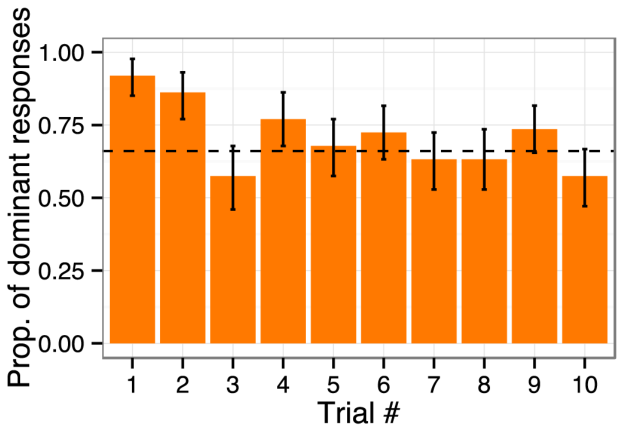

Figure 4: Proportion of majority responses at each successive trial in the 10-trial condition (total: 87 participants). Dashed line indicates participants' estimate of the proportion of orange gumballs (66%). Error bars are bootstrapped 95% confidence intervals.

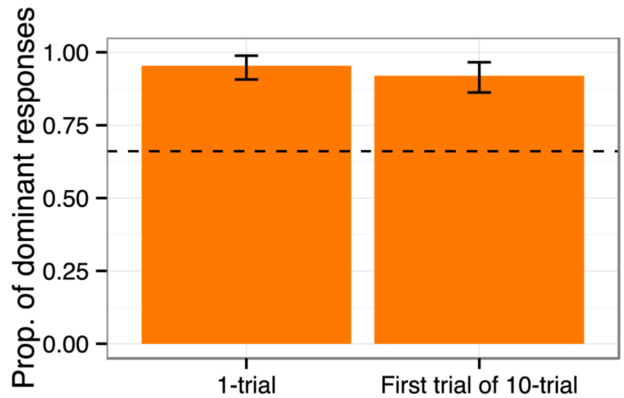

Figure 5: Comparing the mean proportion of majority (orange) responses in the only trial of the 1-trial condition and the first trial of the 10-trial condition. Dashed line indicates participants' estimate of the proportion of orange gumballs (66%). Error bars are bootstrapped 95% confidence intervals.

come out of a gumball machine next. Participants in the multiple-trial group were asked to guess the colors of the next 10 gumballs, while participants in the single-trial group were only asked to guess once. As a population, participants the single-trial condition collectively maximized, while participants in the multiple-trial condition were more likely to probability-match. Interestingly, though, participants in the multiple-trial condition show strong sequence effects: usually favoring the majority option on the first two trials, and the minority option on the third. This creates a representative sample of the population after only three trials. This finding supports our hypothesis—the incremental-representativeness hypothesis—positing the existence of a single mechanism that seeks to create incrementally representative samples at each trial of the experiment; explaining the aggregate patterns of maximizing and probability-matching we observe (in the single-trial and multiple-trial conditions, respectively), as well as the order effects we see in initial responses.

Participants in Experiment 2 also preferred sequences that started with the majority gumball. Especially notable are sequences where the non-ordered probabilities should be identical, such as '110'and '011'. As highlighted in Figure 6, though, both free-responses and ratings over sequences show a significant preference for the order that begins in a more representative manner. Sequences that do not begin with a majority gumball are regarded as less likely. This suggests that Kahneman and Tversky's representativeness heuristic is employed from the very first trial. Participants aim to make every sample—even a sample of one—a faithful representation of the distribution of the population. Further experiments with different population distributions and trial numbers are planned in order to verify this assumption.

Much like many prior studies, our results also show that not every participant uses the same strategy. For example, as in Figure 3, a few participants in the multiple-trial condition chose to maximize, and often justified this strategy. One subject explicitly stated that they thought if they chose the majority choice in every trial, they would "get more correct guesses". Other participants in this condition, while probability-matching and explicitly referring to the probability distribution in their comments, did not display any sequential effects. They chose the majority response consistently for the first fraction of trials, then chose the minority response exclusively for the rest. In spite of the above, we have chosen to report aggregate analyses whenever possible. While an aggregate analysis across participants may gloss over individual idiosyncrasies, it does reveal important patterns of behavior, and point to the causal mechanisms behind them.

Conclusion

While numerous studies have been conducted using binary-choice tasks, with participants' responses varying with a myriad of experimental variables, the mechanisms underlying participants' behaviors on these tasks remain poorly understood. In this study, we propose a unifying mechanism that explains seemingly different behavior when participants are given one vs. more trials, as well as initial order effects we observed when subjects were asked to give multiple responses. We explored these phenomena particularly in the absence of training or feedback. We found that participants given a single trial tend to maximize, while participants given multiple trials probability-match, replicating previous results (Vulkan, 2000; James & Koehler, 2011). However, we found no significant difference between participants' responses in the single-trial condition and those in the first trial of the multiple-trial condition, indicating that maximizing behavior is not exclusively a response to unique gambles. This sug-

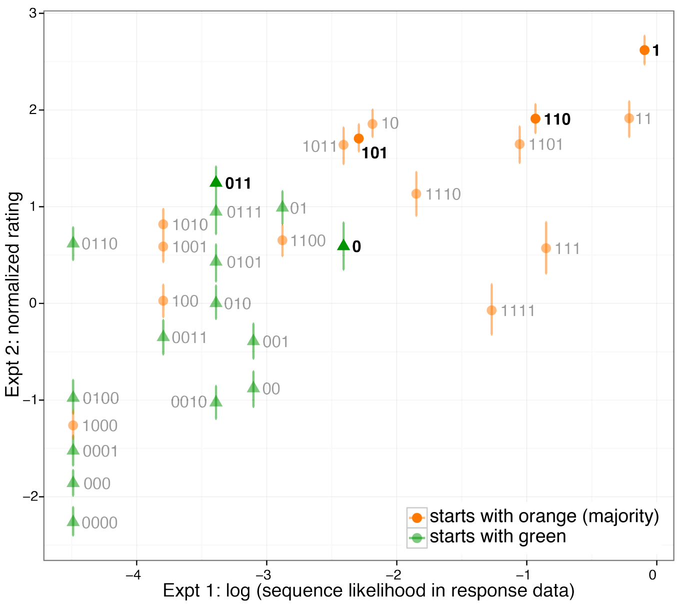

Figure 6: Average normalized rating of each possible sequence (obtained from Experiment 2), versus the likelihood of participants generating this sequence initially (obtained from Experiment 1), in log scale. These two variables are significantly correlated (Pearson correlation coefficient = 0.536 $\pm$ 0.021, $p < 10^{-15}$)

gests a single, underlying mechanism responsible for generating behavior in both of these conditions. While participants seem to probability-match over the course of the entire experiment, they do not do so exclusively for each trial. We also observe sequence effects—particularly in the initial three trials—suggesting that participants attempt to provide samples that are not only representative of the population distribution over the course of the entire experiment, but are also incrementally representative. We call this effect the incremental-representativeness hypothesis—participants seek to provide a sequence of responses that, truncated at any point during the experiment, still produces a sample representative of the population. This is in contrast to a trial-by-trial probability-matching mechanism, which fails to explain maximizing behavior observed in single-trial experiments.

Author Contributions

D.R. developed the study concept. All authors contributed to the study design. Testing and data collection were performed by H.A. H.A. performed the data analysis and interpretation under the supervision of C.K. H.A. drafted the manuscript, and D.R. and C.K. provided critical revisions. All authors approved the final version of the manuscript for submission.

Acknowledgments

We thank the University of Rochester for funding this research. Special thanks to Frank Mollica and Amanda Yung for their assistance in setting up and running the task, and Steve Piantadosi and members of the Kidd Lab for helpful feedback and comments regarding this work.

References

Anderson, Norman H (1960). “Effect of first-order conditional probability in a two-choice learning situation.” In: *Journal of Experimental Psychology* 59.2, pp. 73–93.

Birnbaum, Michael H & Sandra V Wakcher (2002). “Web-based experiments controlled by JavaScript: An example from probability learning”. In: *Behavior Research Methods, Instruments, & Computers* 34.2, pp. 189–199.

Denison, Stephanie & Fei Xu (2014). “The origins of probabilistic inference in human infants”. In: *Cognition* 130.3, pp. 335–347.

Derks, P L & M I Paclisanu (1967). “Simple Strategies in Binary Prediction by Children and Adults.” In: *Journal of Experimental Psychology*.

Edwards, Ward (1961). “Probability learning in 1000 trials.” In: *Journal of Experimental Psychology* 62.4, pp. 385–394.

— (1965). “Optimal strategies for seeking information: Models for statistics, choice reaction times, and human information processing”. In: *Journal of Mathematical Psychology* 2.2, pp. 312–329.

Friedman, Daniel & Dominic W Massaro (1998). “Understanding variability in binary and continuous choice”. In: *Psychonomic bulletin & review* 5.3, pp. 370–389.

Gardner, R Allen (1957). “Probability-Learning with Two and Three Choices”. In: *The American Journal of Psychology* 70.2, p. 174.

Goodnow, Jacqueline Jarrett (1955). “Determinants of Choice-Distribution in Two-Choice Situations”. In: *The American Journal of Psychology* 68.1, p. 106.

Grant, David A, H W Hake & John P Hornseth (1951). “Acquisition and extinction of a verbal conditioned response with differing percentages of reinforcement.” In: *Journal of Experimental Psychology* 42.1, pp. 1–5.

Green, C S et al. (2010). “Alterations in choice behavior by manipulations of world model.” In: *Proceedings of the National Academy of Sciences of the United States of America* 107.37, pp. 16401–16406.

Hake, H W & R Hyman (1953). “Perception of the statistical structure of a random series of binary symbols.” In: *Journal of Experimental Psychology*.

James, Greta & Derek J Koehler (2011). “Banking on a bad bet. Probability matching in risky choice is linked to expectation generation.” In: *Psychological Science* 22.6, pp. 707–711.

Kahneman, Daniel & Amos Tversky (1972). “Subjective Probability: A Judgment of Representativeness”. In: *The Concept of Probability in Psychological Experiments*. Dordrecht: Springer Netherlands, pp. 25–48.

Rubinstein, Irvin (1959). “Some factors in probability matching.” In: *Journal of Experimental Psychology* 57.6, pp. 413–416.

Shanks, David R, Richard J Tunney & John D McCarthy (2002). “A re-examination of probability matching and rational choice”. In: *Journal of Behavioral Decision Making* 15.3, pp. 233–250.

Siegel, Sidney (1961). “Decision Making and Learning under Varying Conditions of Reinforcement”. In: *Annals of the New York Academy of Sciences* 89.5, pp. 766–783.

Siegel, Sidney & Donald Aaron Goldstein (1959). “Decision-making behavior in a two-choice uncertain outcome situation.” In: *Journal of Experimental Psychology* 57.1, pp. 37–42.

Toda, M (2009). “Guessing Sequence under Various Conditions of Payoff”. In: *Japanese Psychological Research*, p. 11.

Vulkan, N (2000). “An economist’s perspective on probability matching”. In: *Journal of economic surveys*.

West, R F & K E Stanovich (2003). “Is probability matching smart? Associations between probabilistic choices and cognitive ability”. In: *Memory & Cognition*.

Xu, Fei & Stephanie Denison (2009). “Statistical inference and sensitivity to sampling in 11-month-old infants”. In: *Cognition* 112.1, pp. 97–104.

Xu, Fei & Vashti Garcia (2008). “Intuitive statistics by 8-month-old infants.” In: *Proceedings of the National Academy of Sciences of the United States of America* 105.13, pp. 5012–5015.