

**Inference from Streaks in Random Outcomes:  
Experimental Evidence on Beliefs in Regime-Shifting and the  
Law of Small Numbers \***

Elena Asparouhova<sup>†</sup> and Michael Lemmon  
David Eccles School of Business  
University of Utah

Michael Hertzel  
W. P. Carey School of Business  
Arizona State University

This version April 2009

---

\*We thank Robert Bloomfield, Alex Edmans, Jeffrey Hales, Spencer Martin, George Wu, an anonymous associate editor and two anonymous referees, the seminar participants at the University of Washington, University of Florida, Arizona State University, the University of Toulouse, the University of Utah, the WFA 2005 meeting, the 2006 EFMA Behavioral Finance Symposium, and the 2006 EFMA Madrid meeting for helpful comments. Financial support was provided by the David Eccles School of Business and National Science Foundation grant SES-0616645 (Asparouhova).

<sup>†</sup>Corresponding author: Elena Asparouhova, Tel: + 1-801-587-3975, Fax: + 1-801-581-3956, E-mail: e.asparouhova@utah.edu, Address: David Eccles School of Business, 1645 East Campus Center Drive, University of Utah, Salt Lake City, UT 84112

## ABSTRACT

Using data generated from laboratory experiments, we test and compare the empirical accuracy of two models that focus on judgment errors associated with processing information from random sequences. We test for regime-shifting beliefs of the type theorized in Barberis, Shleifer, and Vishny (1998) (BSV) and for beliefs in the “law of small numbers” as modeled in Rabin (2002). In our experiments, we show subjects randomly generated sequences of binary outcomes and ask them to provide probability assessments of the direction of the next outcome. Inconsistent with regime-shifting beliefs, we find that subjects are *not* more likely to predict that the current streak will continue the longer the streak. Instead, consistent with Rabin, subjects are more likely to expect a reversal following short streaks and continuation after long streaks. Results of a “test of fit” analysis based on structural estimation of each model also favor the model in Rabin. To provide more insight on Rabin, we use an additional experimental treatment to show that as the perception of the randomness of the outcome-generating process increases, subjects are more likely to predict reversals of current streaks.

# I. Introduction

Experimental research by cognitive psychologists provides a multiplicity of evidence showing that cognitive biases have profound effects on belief formation and revision. Translating the wide variety of documented biases into a unified tractable framework for understanding belief revision has provided an important challenge for decision theorists. One promising approach to synthesizing this evidence is “quasi-Bayesian” where decision makers, because of one or more of the documented cognitive biases, have a mistaken view of the world, but otherwise act as Bayesians in updating beliefs. An important property of this class of models is that rational decision making obtains in the limit as the parameters representing the cognitive biases approach zero, or as the biases disappear. In this paper, we experimentally test and compare the empirical accuracy of two leading quasi-Bayesian models that focus on judgment errors associated with processing information from randomly generated data.

Barberis, Shleifer, and Vishny (1998) (hereafter BSV) develops a model motivated by two well-documented systematic biases that arise when people form beliefs: conservatism and representativeness.<sup>1</sup> These biases are captured in a regime-shifting framework where the (binary) outcome process follows a random walk, but the decision maker (an investor in this case) instead holds the flawed belief that the process switches between a “reversal” regime (in which consecutive outcomes tend to reverse themselves in sign) and a “continuation” regime (in which consecutive outcomes are more likely to be of the same sign). A key implication of the BSV model that we investigate in our empirical tests is that the longer the observed streak of consecutive like outcomes, the higher the probability that the decision maker assigns to the next outcome continuing the streak, i.e., the probability of a streak continuing increases monotonically with streak length.<sup>2</sup>

---

<sup>1</sup>Conservatism refers to the tendency to underweight new evidence relative to prior beliefs and is suggestive of underreaction. Representativeness bias comes in many forms. Relevant here is the belief that even small samples will reflect the properties of the parent population. This can lead to overinference from small samples and is suggestive of overreaction.

<sup>2</sup>As described in BSV (1998, p. 310), “when a positive surprise is followed by another positive surprise, the investor raises the likelihood he is in the trending regime, whereas when a positive surprise is followed by a negative surprise, the investor raises the probability he is in the mean-reverting regime.”

Rabin (2002) provides a model that is based on a form of representativeness bias that Tversky and Kahneman (1971) refer to as the “law of small numbers.” The model assumes a random outcome-generating process and depicts the flaw in reasoning as the decision maker’s false belief that outcomes are, instead, generated by draws from an “urn” of size  $N$  *without replacement*. Two counteracting effects arise when individuals do not know and, thus, must infer the the urn rate. The first effect is the “gambler’s fallacy”—the erroneous belief that future outcomes will “balance” the observed historical sequence towards the presumed rate.<sup>3</sup> The second effect is that, after observing a streak of like outcomes, beliefs about the rate are biased towards the more frequently observed outcome leading to the false expectation that the streak will continue, something commonly referred to as the “hot hand effect.”<sup>4</sup> The interaction of these two effects results in an overestimate of the probability of a short streak reversing (as a short streak has a minimal effect on the rate prior such that the gambler’s fallacy prevails) and that of a long streak continuing (as a long streak leads to an excessive weight on the rate inferred thereby swamping the gambler’s fallacy effect). Thus, in contrast to the BSV model, where the decision maker monotonically increases the probability of continuation as streak length increases, the relation between streak length and expectation of continuation in Rabin is non-monotonic, i.e., in Rabin, the decision maker initially decreases the probability of continuation and then switches to increasing the probability of continuation at longer streak lengths.

Our empirical investigation of the BSV and Rabin models is based on data from laboratory experiments. In the experiments, subjects are presented with randomly generated sequences of eight binary (UP or DOWN) outcomes and then asked to make a probability assessment on the likelihood that the next, ninth, outcome will be UP. We use an incentive compatible mechanism to elicit truthful revelation of subjective probabilities. Subjects are presented with

---

<sup>3</sup>Gambler’s fallacy is sometimes referred to as a manifestation of the “law of small numbers” where individuals believe that even small samples should be representative of the population as a whole; a classic example of this bias is when, after a string of reds at the roulette wheel, bettors expect that a “black is due” in order to balance out the string of reds. A number of studies including those of Tversky and Kahneman (1971), Burns and Corpus (2004), Bar-Hillel and Wagenaar (1991), Rapoport and Budescu (1992), Clotfelter and Cook (1993), Terrell (1994), and Croson and Sundali (2005) present experimental evidence on gambler’s fallacy. For a comprehensive review of the literature, see Rabin (2002).

<sup>4</sup>In comparing hot hand and gambler’s fallacy effects Croson and Sundali (2005) note “The gambler’s fallacy is a belief in negative autocorrelation of a non-autocorrelated random sequence of outcomes. In contrast, the hot hand is a belief in positive autocorrelation of a non-autocorrelated random sequence of outcomes.”

100 such eight-outcome sequences and, thus, are asked to make 100 probability assessments. These probability assessments form the basis for all of our empirical analyses.

In our first set of tests, we focus on the overall ability of each model to explain the data. In this analysis, we use the experimental data to structurally estimate the parameters of each model for each subject. We then test for goodness of fit by computing, and then comparing, the root mean squared error (RMSE) under Rabin, BSV, and two benchmark models. The first benchmark model is the “Correct” model (that always predicts 50% as the likelihood that the next outcome is UP); the second benchmark model is the Bayesian model, in which the decision maker is learning the urn rate but knows that the drawing is made with replacement. We find that both of the behavioral models provide a significantly better fit than the benchmarks, and that the Rabin model provides a significantly better fit than the BSV model. Relative to the Correct model, the Rabin model provides a 14% improvement in fit as compared to only a 6.9% improvement provided by the BSV model. Of the 14% improvement in the Rabin model performance, 6.6% is due to the gambler’s fallacy effect and 7.4% is due to the rational learning of the urn rate, a result that we obtain by comparing the fit of the Rabin model to that of the Bayesian benchmark model.

In order to provide additional evidence and some insight into why the Rabin model provides a better fit of the data, we conduct a simulation analysis to examine how the two models compare in explaining the response of our subjects to streaks in the data. (For this analysis, streaks are measured as the number of like outcomes leading up to the outcome that the subjects are making predictions on. Streaks can range in length from one to eight.) Specifically, for this analysis, we use each subject’s parameter estimates obtained from our structural estimations to simulate (for that subject) the predictions of decision makers that act according to each model. We then compare the simulated predictions from each model to the actual predictions made by our experimental subjects. The results of this analysis favor the Rabin model; i.e., for all streak lengths, the average simulated predictions of the Rabin decision maker are closer (in terms of sign and magnitude) to the actual predictions of our experimental subjects (which are non-monotonic across streak length.) In contrast, the average simulated BSV (and Bayesian) decision makers’ predictions are strictly increasing in streak length.

To more formally characterize the non-monotonic relation between streak length and the subjects’ probability assessments, we conduct a regression analysis where we divide the observed sequences into those that contain “short” and “long” streaks and include both in a piece-wise linear (spline) regression on the subjects’ probability assessments. In addition to the actual experimental data, we perform the above analysis on the simulated BSV, Rabin, and Bayesian data. The results of this analysis are inconsistent with the BSV prediction that individuals are more likely to believe they are in the continuation regime the longer the observed streak; i.e., we find that subjects are *not* more likely to predict that the current streak will continue the longer the streak.<sup>5</sup> The results are also inconsistent with the Bayesian model prediction that the subjects’ probability assessments are strictly increasing in streak length. Instead, consistent with Rabin, when streak length is short (less than or equal to 3) subjects are more likely to expect a reversal the longer the streak. Also, consistent with Rabin, we find that the above effect reverses for longer streaks, i.e., conditional on streak length exceeding 3, subjects are more likely to expect that the streak will continue the longer the streak. Thus, in the aggregate, as allowed by the Rabin model but not by either the BSV or the Bayesian model, the effect of gambler’s fallacy strengthens initially as streak length increases and weakens as the streak becomes longer.

Finally, we also consider an alternative experimental treatment that differs only in that the subjects are told that the underlying process is a random walk as generated by flips of a fair coin. Burns and Corpus (2004) show that subjects are more likely to predict reversals of streaks as their perception of the randomness of the outcome-generating process increases. In the context of Rabin, increasing the perception of randomness increases the impact of the gambler’s fallacy bias and decreases the impact of the hot hand effect. Comparing the results from the two treatments, we find evidence consistent with this prediction.

The two treatments, because of their differing learning environments, also allow us to address the Brav and Heaton (2002) argument that it is difficult, if not impossible, to distinguish between “behavioral” models that rely on individual irrationality and “rational structural un-

---

<sup>5</sup>Ours is not the first study to empirically challenge the assumptions underlying the BSV model. Durham, Hertz and Martin (2005) provide contradictory evidence from the football betting market (that we discuss later in the paper.) Massey and Wu (2005) present a model and experimental evidence suggesting that the way in which agents are quasi-Bayesian is different than that assumed in BSV.

certainty” models, where individuals make decisions rationally, but take time to learn about the structure of the underlying economic environment. In our context, the Brav and Heaton challenge arises from the fact that although a rational decision maker would eventually learn that draws from the urn are made with replacement, his decision-making behavior, while learning, could be indistinguishable from that of a decision maker suffering from the gambler’s fallacy bias. We find that the magnitude of the gambler’s fallacy effect is the same across treatments, which is inconsistent with the hypothesis that what we observe is due to subjects’ learning. Overall, we believe that the evidence we document suggests that the underlying structure of the Rabin model captures important features of decision-making found in the data.

The remainder of the paper proceeds as follows. Section II gives a detailed description of the models of BSV and Rabin. Section III provides our experimental design and the description of the experimental sessions. Our empirical methodology and the statistical results follow in Section IV. Section V concludes with a brief discussion and some thoughts regarding future research.

## II. Theoretical Framework and Empirical Implications

The models in Rabin (2002) and Barberis, Shleifer, and Vishny (1998) are motivated by the question of how people with certain cognitive biases think about sequences of random outcomes. The decision makers in both models use the history of realizations of a binary random variable to infer the distribution of this variable in the next period. The updating process is “quasi-Bayesian” in that the decision makers use Bayes’ rule to update some of the model parameters but are “stubborn” about others and as a result never learn the true data-generating process. The models differ with respect to the nature of the underlying cognitive biases and on how these affect the decision making process. In our empirical analysis, we structurally estimate the models using our experimental data and compare goodness of fit. Accordingly, we provide details of each model below.

## A. BSV model

Motivated by evidence of underreaction and overreaction in financial markets, Barberis, Shleifer, and Vishny (1998) provide a model of investor behavior that is consistent with the financial market findings. Although the BSV theory is cast up as domain-specific to financial markets, the underlying model of decision-making, which draws on well-documented evidence of psychological biases, is potentially applicable in other settings where decision makers are confronted with random data. Specifically, Barberis, Shleifer, and Vishny (1998) model the psychological effects of conservatism and representativeness bias in a regime-shifting framework where the true data-generating process is random, but the decision maker holds the flawed belief that the world shifts between two states (regimes), each of which has a different model governing the data-generating process. The governing models in each state differ only with respect to the probability ( $\pi$ ) that the next outcome will be the same as the previous one. In Model 1, this probability,  $\pi_L$ , is low (less than 0.5) and, thus, outcomes in this regime are mean-reverting. The actions of the decision maker in this state are consistent with conservatism bias, i.e., expecting a reversal, the DM underreacts to the most recent observation. In Model 2, the probability of continuation,  $\pi_H$ , is high (greater than 0.5) and, thus, outcomes in this regime trend. A DM in this world behaves in a manner consistent with representativeness bias; i.e., expecting continuation, the DM overreacts to the most recent observation.

Having observed a sequence of realized outcomes, the BSV decision maker sees her task as attempting to determine which of the two states is governing the outcome-generating process. A crucial assumption of the model is that the DM is dogmatic about the parameters  $\pi_L$  and  $\pi_H$  as well as the parameters ( $\lambda_1$  and  $\lambda_2$ ) that determine the probabilities of switching from the reversal to the continuation regime, and vice versa. (If, instead, the DM were to update these probabilities, she would ultimately learn the true data-generating process.) As a consequence, the longer the streak of like outcomes that the DM observes, the more likely she believes she is in the continuation regime and the higher the probability she assigns to the next outcome continuing the streak, i.e. this probability is monotonically increasing in streak length. The latter is generally not true for the Rabin model, as we describe below.

## B. Rabin Model

The Rabin (2002) model of the “law of small numbers” assumes that people mistakenly believe that the binary outcomes in random sequences are generated from an “urn” of size  $N$  without replacement, although draws are actually made *with* replacement, i.e., the data-generating process is a random walk. To reconcile his belief that the urn is finite with observing very long sequences, the decision maker also believes that the urn is renewed every  $K$  draws ( $K < N$ ). The Rabin DM (called Freddy) is dogmatic about the size of the urn and the frequency with which the urn is renewed. If he were to use the history of realizations to update  $K$  and  $N$ , Freddy would ultimately correctly infer that the urn is either infinitely large or that it is renewed every period (i.e., the draws are made *with* replacement), both leading to correct inference about the distribution of the random variable. When Freddy is uncertain about the urn rate, we use two parameters,  $r$  and  $q$ , to describe his prior about the urn rate distribution. Apart from being “stubborn” about  $N$  and  $K$ , Freddy uses Bayes’ rule for updating his priors on the urn rate.

One implication of the Rabin model is that when Freddy knows that the urn rate is 50% he succumbs to the gambler’s fallacy, i.e., after observing an unbalanced sequence he believes it is more likely that the next outcome will “balance” the sequence toward the rate of 50%. When Freddy does *not* know the urn rate and makes inferences about the rate from past observations, there are two counteracting effects. The first effect is that given a rate, he predicts outcomes in the direction that would balance the observed sequence towards that rate, i.e. the standard gambler’s fallacy effect. The second effect is that after observing a disproportionate number of like outcomes, Freddy infers a rate that is biased towards the more frequent outcome. Generally, the interaction of these two effects leads to a non-monotonic relationship between the length of the streak of like outcomes that Freddy observes and the probability that he assigns to the next outcome continuing the streak.

### III. Experimental Setup

Our experimental design is motivated by a recent study by Bloomfield and Hales (2002) that presents evidence consistent with regime-shifting beliefs of the type envisioned by BSV. In their laboratory experiments, subjects observe eight separate graphical representations of historical sequences (and their mirror images) of eight binary outcomes (UP or DOWN). Participants are told that the sequences are generated from a random walk and a pricing mechanism is used to elicit their expectations about the direction of the ninth outcome in the sequence. Consistent with BSV, Bloomfield and Hales find that subjects consistently rely on the prevalence of past performance reversals when assessing the likelihood of future reversals. More specifically, the subjects showed “a strong tendency to predict reversion after seeing many reversals and to predict trending after seeing few recent reversals.” (p.412.)

We argue that the Bloomfield and Hales (2002) experimental design cannot provide a definitive test of BSV because the set of sequences shown to their subjects are not consistent with what would be observed under a random walk process, but instead are more consistent with what would be expected if the true underlying process was of a regime-shifting type. A simple chi-square goodness of fit test based on the frequency of reversals strongly rejects that the set of sequences used in their experiment were drawn from a random walk process. Furthermore, consistent with an underlying regime-shifting process, the sequences have far too many observations in the tails of the distribution of reversal rates. Although there may be sound methodological reasons for employing extreme sequences, one unintended consequence is that it results in an experiment that is unable to distinguish whether subjects rationally conclude that the underlying process is of a regime-shifting type or whether the subjects’ belief in regime-shifting arises from behavioral biases as suggested by BSV.

We, therefore, modify the Bloomfield and Hales methodology by presenting our subjects with eight-outcome sequences that are each drawn independently from a (discrete) random walk distribution. Then, as in Bloomfield and Hales, we ask participants to report their

probability assessments that the next (ninth) outcome will be UP.<sup>6</sup> We elicit the subjects' probability assessments using an incentive compatible mechanism; namely the variant of the quadratic scoring rule described in Offerman and Sonnemans (2004).<sup>7</sup> Once a probability is reported, the ninth (randomly generated) outcome of the corresponding sequence is revealed, the payoff for that "round" is realized, and a new sequence is shown. Subjects are presented with 100 sequences and, thus, are asked to make 100 separate probability assessments. We vary the order and the direction in which the patterns are presented to subjects as follows. There are four groups of subjects in each session. Group 1 observes the 100 randomly generated sequences. Group 2 observes the same sequences, however, each UP (DOWN) outcome is replaced by a DOWN (UP) outcome. Group 3 observes Group 1's sequences but in reverse order. Group 4 observes Group 2's sequences but in reverse order.

We further modify the Bloomfield and Hales setup by *not* telling the subjects that outcomes are generated by a random walk process. Our main treatment, Treatment 1, is designed to satisfy the information condition of both BSV and Rabin that the DM is not informed about the outcome-generating process. In this treatment, we inform the subjects that they are observing sequences of performance surprises (that can be good or bad, as represented by UP or DOWN outcomes) of the same firm.<sup>8</sup> The information on how the sequences are generated is, however, withheld from the subjects.<sup>9</sup>

As discussed in Section 2, Rabin makes divergent predictions based on whether or not the DM knows the data-generating mechanism. Consequently, to further investigate the model,

---

<sup>6</sup>A snapshot of the experimental software displaying a single eight-outcome sequence is presented in Figure 1. The software eTradeLab used in our experiments is web-based and the graphical application we use was designed specifically for our experiment.

<sup>7</sup>Offerman and Sonnemans (2001) study the properties of this particular quadratic scoring rule and find that "the reported mean absolute difference between true subjective and reported probabilities is for most subjects smaller than a few percentage points. Deviations from truth telling are not systematic. The scoring rule does not bias the results."

<sup>8</sup>Although we propose that BSV is a more generalizable "domain-general" model (the Rabin model is "domain-general" by construct), an interesting question is the extent to which the economic setting affects the subjects' behavior. In the context of our study, a potential future experiment might explore whether changing from "up/down" stimuli to any other binary stimuli (for example, stock returns) changes the inferences that subjects draw.

<sup>9</sup>The instructions for this treatment with the exact language used to describe the data-generating process to the participants are provided in the E-companion for this paper on the Management Science web page. It can also be found at [http://leef.business.utah.edu/ms2/frames\\_ms.html](http://leef.business.utah.edu/ms2/frames_ms.html). The Excel macro used to explain the quadratic scoring rule can be found at <http://leef.business.utah.edu/ms2/ProbabilityAssessment.xls>.

we employ an additional treatment that differs from Treatment 1 in that the subjects are told what the outcome-generating process is. In Treatment 2, we inform the subjects that they are observing outcomes generated by the flips of a fair coin (with no reference to firm performance surprises).<sup>10</sup>

A total of 92 participants took part in the experimental sessions, with 46 participants in each treatment. The sessions were conducted in the Fall of 2007. Approximately half of the participants in each treatment were from the University of Utah; the other half were from Arizona State University. Each experimental session lasted about one hour. The average payoff per subject was \$24. Upon arrival at the laboratory, subjects were seated in front of computer terminals. They were instructed to go to a web page that contained the experimental instructions and the links to the practice and the actual sessions. One of the experimenters read aloud the instructions. The subjects were advised that they could ask questions, but that they should do so privately (by raising their hand and having one of the experimenters come to them to answer their question(s)). An Excel macro was designed to explain that the expected payoff would be maximized by truthful reporting of subjective probability assessments. Using the macro, the subjects could explore the quadratic scoring rule that determined their payoffs. Each subject could take as long as s/he needed to complete the training. To ensure that subjects understood the instructions they were asked to fill out a short questionnaire and could not proceed to the practice rounds until they answered the questions correctly. In the practice rounds, the subjects observed two sequences (same for all subjects), made predictions, observed the payoffs, and, in general, familiarized themselves with the experimental software. After completing the practice sessions, subjects completed the actual experiment at their own pace.

## IV. Empirical Analysis

The basic unit of analysis in our empirical investigation is a subject's probability assessment, after observing a sequence of eight outcomes, that the next (ninth) outcome is UP. Whether

---

<sup>10</sup>The URL for Treatment 2 is [http://leef.business.utah.edu/ms4/frames\\_ms.html](http://leef.business.utah.edu/ms4/frames_ms.html).

this assessment translates into an expectation of continuation or reversal depends, of course, on the last (eighth) outcome of the observed sequence. To account for this, we use the Bloomfield and Hales (2002) “signed reaction measure.” To calculate this measure, we first subtract 0.5 from each of the participant’s 100 probability assessments; this “deviation” ranges between -0.5 and 0.5, with positive numbers indicating that the subject places a higher likelihood on the next outcome being UP, and negative numbers indicating that the subject places a higher likelihood on the next outcome being DOWN. To obtain the signed reaction measures, we then multiply this deviation by -1 if the last outcome is DOWN. Thus, we obtain a measure where expectations of continuations are positive and expectations of reversals are negative.

Table I presents summary information (mean, median, and interquartile range) for our sample of 9200 probability assessments (4600 each for Treatments 1 and 2) and for the associated sample of signed reaction measures. The results for both treatments show a slight bias in predicting an UP movement. For Treatment 1 the median (mean) reported probability of an UP movement is 0.50 (0.52), with an interquartile range of [0.40, 0.65]. The results for Treatment 2 are nearly identical with a median (mean) reported probability of 0.5 (0.52) and an interquartile range of [0.40, 0.70]. (We conjecture that the positive connotation of the UP stimulus may have caused subjects to place higher subjective probability on this event.)

In both treatments the sample median (mean) signed reaction measure is 0 (0.006). The interquartile ranges are  $[-0.1, 0.13]$  and  $[-0.14, 0.15]$  for Treatments 1 and 2 respectively. These results suggest that unconditionally there is no evident tendency towards continuation or reversal in subjects’ behavior. We note that the ability of the BSV model to explain underreaction to recent performance relies on an assumption that the DM’s overall tendency is to predict reversals. However, as noted in Section III, we do not impose this condition in our estimation, and as such this finding has no bearing on the goodness of fit tests we perform on the BSV model.

In the next subsection we use the raw probability assessments to structurally estimate the models and perform a goodness of fit test of their relative performance. We then provide additional descriptive analysis by reporting the summary information presented in Table I for the eight subsamples of sequences defined by the lengths of the sequence-ending streaks. Guided

by the observations from this simple analysis, we then use the signed reaction measures in a regression analysis that tests the differing implications of the models regarding how individuals respond to streaks in performance.

## A. BSV vs. Rabin (Treatment 1)

### A.1. Test of Fit Analysis

We begin our analysis of the experimental data by structurally estimating the parameters of the BSV and Rabin models. This analysis provides a direct test of the ability of the two models to describe the data. The parameters of both models are estimated at the subject level using a standard non-linear least squares procedure.<sup>11</sup> We compare the two models against each other and against two benchmark models using root mean squared error (RMSE) as a measure of goodness of fit.

Starting with BSV, for each subject ( $i$ ), we estimate the four parameters of interest, namely,  $\pi_{L,i}$  (the probability of continuation in the reversal regime),  $\pi_{H,i}$  (the probability of continuation in the continuation regime),  $\lambda_{1,i}$  (the probability that the regime switches from reversal to continuation), and  $\lambda_{2,i}$  (the probability that the regime switches from continuation to reversal). To deliver the desired empirical implications of short-run underreaction and long-run overreaction, the BSV model imposes a constraint on the relationship between the above parameters (see Proposition 2 of Barberis, Shleifer, and Vishny (1998)). Since this constraint only concerns the empirical implications of the model we do not impose it in our estimation. The objective function in the minimization algorithm for BSV is

$$F(\pi_{L,i}, \pi_{H,i}, \lambda_{1,i}, \lambda_{2,i}) = \sum_{s=1}^{100} (P_i^s - Pr(y_9^s = UP | y_1^s, \dots, y_8^s, \pi_{L,i}, \pi_{H,i}, \lambda_{1,i}, \lambda_{2,i}))^2,$$

where  $P_i^s$  is subject  $i$ 's reported probability of an UP outcome after seeing sequence  $s$ ,

$y_j^s, j=1, \dots, 8$ , are the observed eight outcomes in sequence  $s$ , and

---

<sup>11</sup>To implement the non-linear least squares estimation, we use a standard algorithm in Matlab, called "fmincon," for finding the minimum of a constrained non-linear multivariable function.

$Pr(y_9^s = UP|y_1^s, \dots, y_8^s, \pi_{L,i}, \pi_{H,i}, \lambda_{1,i}, \lambda_{2,i})$  is the conditional probability that a (BSV) decision maker with parameters  $(\pi_{L,i}, \pi_{H,i}, \lambda_{1,i}, \lambda_{2,i})$  assigns to the ninth outcome in sequence  $s$  being UP. The formula for this posterior probability is provided in the Appendix.

For the Rabin model, for each subject, we estimate the urn size  $N_i$ , the renewal rate  $K_i$ ,<sup>12</sup> and two additional parameters  $q_i$  and  $r_i$  that characterize the subject's prior about the possible urn rates. We assume that the possible values of the urn rates are

$$\{10\%, 20\%, 30\%, 40\%, 50\%, 60\%, 70\%, 80\%, 90\%\}$$

and that the corresponding prior probabilities associated with each of these values are

$$\Pi_i = \{q_i, q_i, 2q_i, 2q_i, 1 - 6q_i - 6r_i, 2r_i, 2r_i, r_i, r_i\},$$

where  $q_i, r_i \geq 0, q_i + r_i \leq 1/6$ .<sup>13</sup>

The objective function for Rabin in the minimization algorithm is

$$G(N_i, K_i, q_i, r_i) = \sum_{s=1}^{100} (P_i^s - Pr(y_9^s = UP|y_1^s, \dots, y_8^s, N_i, K_i, q_i, r_i))^2,$$

where  $P_i^s$  is subject  $i$ 's reported probability after sequence  $s$ ,

$y_j^s, j=1, \dots, 8$ , are the observed eight outcomes in sequence  $s$ , and

$Pr(y_9^s = UP|y_1^s, \dots, y_8^s, N_i, K_i, q_i, r_i)$  is the conditional probability that a (Rabin) decision maker with parameters  $(N_i, K_i, q_i, r_i)$  assigns to the ninth outcome in sequence  $s$  being UP. The formula for this posterior probability is provided in the Appendix.

We compare the fit of the BSV and Rabin models against each other and against two rational benchmarks. The first benchmark, which we refer to as the ‘‘Correct model,’’ is where

<sup>12</sup>As a robustness check, we repeat the estimation fixing  $K_i$  to be 9 for all  $i$  (results not reported). Excluding the urn renewal parameter results in only a small loss in fit (the median RMSE decreases by less than 0.5%).

<sup>13</sup>The first sequence that subjects observe contains both UP and DOWN outcomes. Thus, subjects know that the rate is not 0% or 100%. For robustness, we also estimated the symmetric prior of  $\{r_i, r_i, q_i, q_i, 1 - 4q_i - 4r_i, q_i, q_i, r_i, r_i\}$ , where  $q_i, r_i \geq 0, q_i + r_i \leq 1/4$ . None of the results that we report below changes when this prior parameterization is used.

the DM always predicts 50% as the probability of the next outcome being UP. The second benchmark is a fully Bayesian model where the DM learns the urn rate (using Bayes' rule) but does not suffer from the gambler's fallacy bias. The Bayesian benchmark model is obtained as a particular case of the Rabin model as the urn size approaches infinity or as the urn renewal rate approaches one. For this benchmark, we estimate the parameters of the prior distribution  $q_i$  and  $r_i$  fixing the values  $N=10,000$  and  $K=1$ . We refer to this second benchmark model as the "Bayesian model."

The structural estimation results are reported in Table II. For each model, we report the median and mean values of the estimated parameters across subjects and the resulting median and mean values of the root mean squared error (RMSE). The results for the BSV and Rabin models are reported in Panels A and B respectively. Panel C reports the results for the two benchmark models.

For the BSV model, the median parameter estimates indicate that, within the continuation regime, subjects place a relatively high probability on continuation (0.59). In contrast, within the reversal regime the likelihood of a reversal (0.53) is only slightly greater than 0.5. However, the likelihood of switching to the reversal regime is quite high (0.24) compared to the likelihood of switching from the reversal regime to the continuation regime (0.01). The mean values of the parameters yield similar inferences. With these parameter estimates, the BSV DM starts by predicting reversals at short streaks and then, as the streak length increases, places higher likelihood on continuation.

For the Rabin model, the median estimated urn size ( $N$ ) is 20 (we do not report the mean urn size because the estimate of  $N$  reaches its upper limit in the least squares procedure for subjects who state a probability of 0.5 for all, or most, of the observed sequences), and the median urn renewal rate ( $K$ ) is 9. Taken together, these estimates of  $N$  and  $K$  indicate that the median subject exhibits the gambler's fallacy bias.

The estimates of the subjects' prior distribution of urn rates provide evidence that the subjects place some weight on the possibility that the urn rate is not equal to 50%. Specifically, based on the individual estimates of  $q$  and  $r$ , the median subject believes that the prior proba-

bility of a 50% urn rate is 0.81 (i.e.,  $0.81 = \text{median}_i(1 - 6q_i - 6r_i)$ , where  $i$  is the  $i$ -th subject,  $i = 1, \dots, 46$ ). Consistent with the tendency to predict an UP movement as documented earlier in Table I, subjects appear to place higher prior probabilities on urn rates above 50% compared to urn rates below 50% (i.e.,  $q < r$ , with both the paired Student’s t-test and the Wilcoxon Signed Rank test rejecting the equality of the means and the medians of the distributions of the two parameters).

Finally, the mean estimates for the fully Bayesian model also indicate that subjects place some, albeit small, prior weight on urn rates other than 50%; the median weight placed on an urn rate of 50% is equal to 0.93. Compared to the Rabin model, the relatively high estimate that the median subject in the Bayesian model (which is nested in the Rabin model) places on the urn rate of 50% is due to the inability of the model to capture the gambler’s fallacy effect that our subjects apparently display. To compensate, the model limits the updating that the decision maker has to do.<sup>14</sup>

Turning to the RMSE results, we first note that the Correct (random walk) model produces a median (mean) RMSE of 0.2077 (0.1906), while the corresponding value for the Bayesian benchmark is 0.1924 (0.1791). Both of the behavioral models improve the overall fit relative to the random walk benchmark: the BSV model produces a 6.9% (7.7%) improvement whereas the Rabin model produces an improvement of 14% (14.1%). By comparing the fit of the Rabin model to that of the Bayesian benchmark, we find that 6.6% (8%), or roughly half of the improvement relative to the random walk model, is due to the gambler’s fallacy bias while the rest can be attributed to rational learning of the urn rate.

The Wilcoxon Signed Rank test (paired Student’s t-test) for equality of the medians (means) of the RMSE distributions resulting from the above estimation indicates that the RMSE for the Rabin model is significantly smaller than that of the BSV model with a p-value

---

<sup>14</sup>Although we do not report the results here, we structurally estimated several versions of the Mullainathan (2002) model where the DM is rational in all other aspects but uses “coarse categories to make inferences,” i.e., after observing the past outcomes, instead of updating continuously using Bayes’ rule, the DM has to choose only one rate from the small set of rates that he thinks are possible and base his decisions on that rate (for example, one set of possible urn rates is  $\{0\%, 10\%, 20\%, 30\%, 40\%, 50\%, 60\%, 70\%, 80\%, 90\%, 100\%\}$ ). Given the set of rates the DM can choose from, the structural estimation focuses on the DM’s prior over those rates. In search of a good fit of the data, we used several parameterizations of the DM’s prior. We also varied the set of available urn rates, or “categories,” but the resulting RMSE’s were always larger than the RMSE of the Rabin model.

(two-tailed) of 0.0025 (0.009).<sup>15</sup> Based on the test of fit analysis, we conclude that the Rabin model is better able to explain the behavior of our subjects.

## A.2. Streak Length Analysis

To offer some insight into why the Rabin model provides a better fit to the experimental data, we investigate how our subjects' responses to streaks of different length compare to the responses predicted by the competing models. Our focus on streaks is driven by (i) the aim of the two models to explain empirical regularities that directly relate to streaks in the data and (ii) the differential predictions that the models make with regard to the DM's behavior after short streaks of like outcomes. In BSV, the DM increases the probability of the next outcome continuing the streak the longer the observed streak, i.e. the probability of continuation is monotonically increasing in streak length. In contrast, in the Rabin model the interaction of the gambler's fallacy bias and the urn-updating effect generally leads Freddy to exaggerate the probability of a short streak reversing and of a long streak continuing. This difference in the predictions of BSV and Rabin is highlighted in Rabin and Vayanos (2007) as follows: "Even in settings where Freddy's error patterns resemble those in BSV, there are important differences. For example, within the set of short streaks, Freddy's expectation of a reversal can increase with streak length, while in BSV it unambiguously decreases."

Consequently, the analysis in this section examines the relation between streak length and the signed reaction measure of the subjects' assessment of the likelihood of continuation, where streak length is defined as the number of like outcomes leading up to the outcome that the subjects are predicting (i.e., the "sequence-ending streak"). As in the previous section, we compare the results of the Rabin and BSV models to the benchmark Bayesian model (we exclude the Correct model in this section as it always predicts a signed reaction measure of 0 independent of the streak length).

---

<sup>15</sup>Both the parametric and the non-parametric tests indicate that the two quasi-Bayesian models have smaller RMSE than either of the benchmarks. The corresponding p-values in those comparison tests for the Rabin model are both smaller than 0.01. The paired Student's t-test cannot reject equality of the mean RMSEs of the BSV and the Bayesian model, with a p-value of 0.09. The Wilcoxon Signed Rank test rejects that the two RMSE medians are equal in favor of the hypothesis that the BSV RMSE is lower, with p-value of 0.01.

**Simulation Analysis** To begin, we first investigate how the two models fare in explaining the responses of our subjects to sequence-ending streaks. Specifically, we compare the relation between the average signed reaction measure and streak length that we observe in our experimental data to the relations obtained when we simulate the predictions of decision makers that respectively act according to each model. The simulated predictions for each subject (according to the BSV, Rabin, and Bayesian models) are derived using that subject’s estimated parameters obtained from the previous structural estimation of each model. We then convert the simulated predictions into simulated signed reaction measures in the same way as we do for the actual predictions from the experimental data.

Table III reports the average signed reaction measure by streak length for the experimental and simulated data. Panel A displays the results for the experimental data. The average signed reaction measure is positive for streaks with length of one, indicating that subjects put slightly more weight on continuation compared to reversal. In contrast, the average signed reaction measure for streak lengths between two and four is negative and is increasing in absolute magnitude with the length of the streak. This is consistent with the gambler’s fallacy bias. For streak lengths between five and eight, the average signed reaction measure is positive, indicating that subjects put more weight on continuation of long streaks. Figure 2(a) provides a visual representation of the experimental data.

Panel B presents the Rabin simulation results. The simulated signed reaction measures exhibit a pattern that is remarkably similar (both in terms of sign and magnitude) to the experimental data. In the context of the Rabin model, the simulated results show that at short streaks the gambler’s fallacy effect dominates, such that reversal is more likely than continuation. Moreover, the probability of reversal increases as the streak length initially increases. As the observed streak reaches a length of five, the rate-updating effect starts dominating and continuation becomes more likely than reversal. Figure 2(b) displays the graph corresponding to these results. The one notable difference between the simulated and the actual data is observed at a steak length of one; the Rabin model cannot explain the continuation observed in the experimental data.

Panel C presents the results obtained from the BSV simulations. Figure 2(c) displays the corresponding graph. As can easily be seen, the BSV model fares worse than the Rabin model in capturing the observed pattern in the experimental data. While the BSV simulation captures the switch from predicting reversal of short streaks to predicting continuation of long streaks, it does not capture the gambler’s fallacy effect at short streaks that we observe in the subject data. Instead, the average simulated signed reaction measures for the BSV model are monotonically increasing in streak length. As the graph in Figure 2(a) clearly shows, the empirical relation between streak length and the signed reaction measure for our experimental subject is *not* monotonic.

The Rabin model nests the Bayesian model as a special case where the decision maker knows that either the size of the urn is infinite or that it is renewed after each draw. A decision maker acting according to the Bayesian model will always predict in favor of the outcome that is of higher frequency in the observed data. As presented in Panel D of Table III (and visually in Figure 2(d)), the simulated average signed reaction measure from the Bayesian model displays monotonicity in the streak length, and no reversal at any streak length. The failure of the Bayesian simulation to display non-monotonicity and reversal helps to explain its inferior performance in the structural estimation comparisons, and its marginally worse performance than the BSV model (which allows for predictions of reversals after short streaks).

**Regression Analysis** In order to formally characterize the non-monotonic nature of the relation between streak length and the signed reaction measure, we follow an approach similar to Durham, Hertz, and Martin (2005) and divide the observed sequences into those exhibiting “short” (*STREAK1*) and “long” (*STREAK2*) streaks and investigate how the reaction measure is related to the observed streak length.<sup>16</sup> In particular, if  $C$  is the cutoff point between short and long streaks, and  $STREAK$  is the length of the sequence-ending streak, we define

---

<sup>16</sup>Using data from the college football point-spread betting market, Durham, Hertz, and Martin (2005) investigate whether the Bloomfield and Hales (2002) laboratory results supporting the BSV model carry over to the marketplace. In one set of tests, they examine whether streaks in outcomes against the spread (measured over the most recent eight games) affect changes in the spread during the week leading up to the next (ninth) game. We discuss their findings and the consequences for BSV and Rabin below.

$$STREAK1 = \begin{cases} STREAK & \text{if } STREAK \leq C \\ C & \text{if } STREAK > C, \end{cases}$$

and

$$STREAK2 = \begin{cases} STREAK - C & \text{if } STREAK > C \\ 0 & \text{if } STREAK \leq C. \end{cases}$$

Both variables are included in a piecewise linear regression (spline regression) that allows us to examine the impact of short and long streaks on expectations of continuation. The estimated model (with subject fixed effects) is

$$Y_{is} = \mu_i + \beta_1 STREAK1_{is} + \beta_2 STREAK2_{is} + \epsilon_{is},$$

$i = 1, \dots, I$ ,  $s = 1, \dots, S$ , where  $I$  is the number of subjects,  $S$  (which equals 100) is the number of observations per subject, and  $Y_{is}$  is the reaction measure of subject  $i$  for observation  $s$ .

The results are presented in Table III for cutoff values of  $C=3$  and 4.<sup>17</sup> The first column presents the results from the experimental data. In both regressions, the F-statistics are statistically significant although the R-squared values of the regressions are relatively low (around 6.5%). For both cutoffs, the estimated coefficients for short streaks are negative; the coefficient is statistically significant for the cutoff at  $C=3$ . The negative coefficient estimates indicate that the longer the streak, conditional on the streak being short, the higher the likelihood that subjects place on reversal. Focusing on the results for  $C=3$ , the magnitude of the coefficient estimate indicates that subjects on average reduce their probability estimate of continuation by approximately 0.9% for each unit increase in streak length for streaks of length up to three.

The analysis of the simulated Rabin data, presented in the second column of the table, also produces a negative STREAK1 coefficient, with a remarkably similar magnitude to that found

---

<sup>17</sup>These cutoff values provided the highest R-squareds among the regressions with cutoff points of  $C=2, 3, 4, 5$ , and 6.

in the experimental data. In contrast, we find a *positive* STREAK1 coefficient when analyzing the simulated BSV and Bayesian data (last two columns). Again, focusing on  $C=3$ , these coefficient estimates imply that for each unit increase in streak length, subjects on average *increase* the probability of continuation by 1.9% for BSV and 1% for the Bayesian model. As shown in the table, our conclusions based on the above comparisons are robust to the cutoff choice and hold for  $C=4$  as well.<sup>18</sup>

Turning to long streaks (where the definition of “long” depends on the cutoff  $C$ ), the results are consistent with all three models. Conditional on the streak being long, the longer the streak the higher the probability that subjects assign to the next outcome continuing the streak (all of the coefficient estimates are statistically significant). For the case of  $C=3$ , the magnitude of the coefficient estimate on STREAK2 from the experimental data indicates that subjects increase their probability assessment by about 1.75% for each unit increase in streak length for long streaks. The coefficient estimates on STREAK2 from the simulated data from the Rabin, BSV, and Bayesian models are all positive and very close in magnitude to the experimental estimate.

The results of the above regression analysis for  $C = 3$  are presented graphically in Figure 3. The figure makes clear why the Rabin model provides a better overall fit to the data.<sup>19</sup> Specifically, unlike the Rabin model, neither the BSV nor the Bayesian model can capture the gambler’s fallacy effect that appears in the experimental subject data for short streaks.

---

<sup>18</sup>Although the evidence in Durham, Hertz, and Martin (2005) (hereafter DSM) is clearly inconsistent with the BSV model (in that bettors appear to have a non-monotonic response to streaks) it is unclear whether it is supportive or not of Rabin. The difficulty of comparing with our findings is that the dependent variable in DHM is the *change in spread* over the week leading up to the game. It is hard to know, for example, whether an increase in the spread following a short streak is due to the arrival of sentiment bettors who believe that the streak will continue or, alternatively, to a market correction of sentiment at the beginning of the week that the streak will reverse. Given the evidence in DHM that closing spreads follow a random walk, we think the second alternative (which is consistent with our finding) is a reasonable interpretation.

<sup>19</sup>In results not presented here we subject the Rabin model to further scrutiny. Because the Rabin decision maker uses the imbalances in the observed sequence to form predictions about the next outcome (and the streak predictions are a consequence of the high correlation between streak length and imbalance), we investigated piecewise linear regression specifications where imbalance was included as an explanatory variable. The coefficients from this regression performed on the Rabin data are again very similar to those from the regression performed on the experimental data. The exact results are provided in the E-companion to this paper.

## B. Treatment Analysis

As discussed earlier, the Rabin model encompasses two informational conditions that produce qualitatively different implications about decision maker behavior. The first condition, which we focus on above, is where the DM does not know the urn rate and learns about it from past outcomes. The second informational condition is the limiting case, where the DM *knows* that the rate is 50%. Under this condition, the DM always predicts reversals, and expects that they are more likely the longer the observed streak. An implication of Rabin is that as the DM becomes more and more confident that the urn rate is 50%, his prior on the distribution of rates approaches this limiting case; in other words, the gambler’s fallacy effect becomes stronger and the urn-updating effect becomes weaker. To investigate, we modify the first experimental treatment by telling the subjects that the observed outcomes are generated by a random walk process as can be represented by a fair coin toss.

We begin by structurally estimating the Rabin model using the experimental data from the modified treatment (Treatment 2). The results are reported in Table V. Consistent with the prediction of Rabin, and our experimental modification, the table shows an increased perception of randomness in the new treatment as indicated by the greater prior probabilities that the subjects place on the urn rate of 50%. Specifically, the table shows that the estimated values of Rabin model parameters  $q$  and  $r$  decline relative to the first treatment; the median parameter estimates indicate that the median (mean) subject places 96% (81%) weight on the urn rate being 50%, as compared to 81% (71%) for the median (mean) subject in Treatment 1.<sup>20</sup> We also note that the median estimates of both the urn size  $N$  and the urn renewal rate  $K$  remain as in Treatment 1;  $N=20$  and  $K=9$ . This finding indicates that the gambler’s fallacy effect remains unchanged across treatments. As we will discuss below, this result can be interpreted as providing additional support for the Rabin model.

---

<sup>20</sup>Work by Fox and Rottenstreich (2003) and Bruine de Bruin et al. (2002), showing that individuals may incorrectly rely on a 50% rate, suggests that it is difficult to draw firm conclusions about rationality by examining the weight that subjects place on an urn rate of 50%. Although in both treatments we focus on the case where the true urn rate is 50%, investigating situations where the data is generated from urns with different rates could provide additional insight into whether subjects incorrectly place too much weight on 50%.

In parallel to our analysis of Treatment 1, we compare the fit of the Rabin models against the two rational benchmarks.<sup>21</sup> The Rabin model produces improvements of 9% (8.4%) relative to the Correct model benchmark as illustrated by the comparison of the median (mean) RMSEs of the two models. By comparing the fit of the Rabin model to that of the Bayesian benchmark, we find that 8% (6.6%) of the improvement relative to the random walk model, is due to the gambler’s fallacy bias while 1% (1.8%) can be attributed to rational learning of the urn rate. This finding stands in sharp contrast to evidence from Treatment 1 that showed rational learning accounting for approximately half of the improvement of the Rabin model over the Correct model. This result is consistent with the subjects’ increased confidence that the urn rate is 50%.<sup>22</sup>

As noted above, despite being told that the sequences are generated by a fair coin toss, the structural estimation results for Treatment 2 indicate that the subjects do not fully believe the experimenters’ assertion that the urn rate is 50%. Thus, we expect the subjects to use the information from the observed sequences to update the urn rate, albeit to a lesser extent than the subjects in Treatment 1. Similarly, because the subjects in Treatment 2 are more confident that the urn rate is 50%, we expect to see a stronger tendency to predict reversals at any streak length. We present evidence in Table VI consistent with this expectation. The table provides a comparison across the two treatments of average signed reaction measures by streak length. A Wilcoxon Signed Rank test (paired Student’s t-test) formally rejects equality across treatments of the eight signed reaction medians (means) in favor of the alternative that subjects in Treatment 2 predict more reversals; the one-tailed  $p$ -value is 0.027 (0.018). Although the tendency to predict reversals gets stronger in Treatment 2, we note continuation following streak lengths of one and two in this treatment. As noted previously, this feature of the data is inconsistent with the Rabin model.

We also estimate spline regressions of the relation between streak length and the subjects’ probability assessments of the likelihood of continuation. The results for cutoff values of  $C=4$

---

<sup>21</sup>In results not reported here we also perform the estimation for the BSV model. Its performance is comparable to that of the Bayesian benchmark model and is significantly worse than that of the Rabin model.

<sup>22</sup>For completeness, we compare the performance of the Rabin model across treatment conditions. Although the difference between the mean RMSEs across treatments is not statistically significant ( $p$ -value=0.148), it shows deterioration of performance in Treatment 2.

and 5 (the specifications with the highest  $R$ -squareds) are reported in Table VII. The graphical representation for the latter specification is depicted in Figure 4. Two findings are of interest. First, the regression coefficients on both short and long streaks for both the experimental and the Rabin simulated data are all more negative than those obtained in the first treatment (see Table IV). For example, the coefficient estimates for STREAK 1 (short streaks) at  $C=4$ , are around three times larger in absolute value than the Treatment 1 estimates, indicating a stronger tendency to predict reversals for short streaks. Second, the cutoff points (indicating where urn updating effects begin to dominate) are larger for Treatment 2 ( $C=4$  or 5) relative to Treatment 1 ( $C=3$  or 4). These findings are both consistent with the Rabin prediction that increased perception of randomness increases the gambler’s fallacy effect. Taken together with the other results in this section, the Treatment 2 findings provide additional evidence that the gamblers’ fallacy effect is an important feature of the data that is captured by the Rabin model, but not by the model in BSV.

Finally, Treatment 2 allows us to address concerns about the difficulty of distinguishing between the behavior of a Freddy (who is dogmatic about the urn size  $N$  and the urn renewal rate  $K$ ) and a rational individual who is learning about  $N$  or  $K$  and as a result exhibits Freddy-like behavior along the way (Brav and Heaton (2002)).<sup>23</sup> Since subjects in Treatment 2 are told the true process, and thus have little to learn about the urn renewal rate or the size of the urn over time, we would expect that a rational decision maker would more quickly learn that drawing from the urn is with replacement, and we would expect to see estimates of the urn size that are significantly larger than those in Treatment 1 (or estimates of  $K$  that are close to 1). This is not the case. The median estimated urn sizes (urn renewal rates) in both treatments are 20 (9), indicating a significant gambler’s fallacy effect that is the same across treatments. This result suggests that what we observe is not due to subjects’ learning and that the gambler’s fallacy is indeed a behavioral “bias” as it manifests itself even in settings in which agents should believe that there is little to learn.

---

<sup>23</sup>The DM will make correct inferences if he learns that  $N = \infty$  or  $K = 1$ .

## V. Conclusion

Cognitive psychologists have amassed considerable evidence that a wide variety of cognitive biases affect judgment and decision-making. A promising class of behavioral models that serves to synthesize the effects of these various biases in a tractable way centers on the decision-making of quasi-Bayesian agents. In this paper, we test and compare the empirical accuracy of two leading quasi-Bayesian models that focus on judgment errors associated with processing information from random sequences. In particular, we examine whether experimental subjects exhibit behavior that is more consistent with regime-shifting beliefs as hypothesized by Barberis, Shleifer, and Vishny (1998) or with beliefs in the law of small numbers as modeled by Rabin (2002). Our results provide evidence supportive of the model in Rabin (2002). We find little evidence consistent with investor belief in regime-shifting of the type envisioned by BSV.

In our first set of tests, we use our experimental data to structurally estimate the parameters of each model and test for goodness of fit against each other and against two benchmark models. We find that both behavioral models provide a significantly better fit than the benchmarks, and that the Rabin model provides a significantly better fit than the model in BSV. For each model, we then use the individual structural parameter estimates to simulate the predictions of decision makers faced with the same task as our subjects and compare outcomes. This comparison analysis, as well as a regression analysis that examines how subjects respond to streaks in the data, yields results that are consistent with the implications of Rabin but not of BSV. Finally, we provide additional insight and evidence consistent with the Rabin model, by altering the information environment in a second set of experiments, and comparing results.

The models we investigate here have natural applications to understanding investor behavior, but have broader applicability to any situation where individuals make decisions when confronted with random data. For example, decision makers that observe forecast errors (e.g., from sales forecasts) may be subject to the same set of influences and exhibit similar behaviors to that envisioned by BSV and Rabin. Further, the models and our findings may serve as useful benchmarks in providing a better understanding of how individuals respond to patterns

in data that actually do contain information. To that end, we hope our analysis encourages other laboratory experiments and that our results lead to better models.

## Appendix

### A. BSV Model

In this section we present the BSV model as it applies to our experimental setup. The decision maker observes a sequence of outcomes  $y_1, \dots, y_t$ ,  $y_t \in \{UP, DOWN\}$ , each of which is generated by one of two models (regimes). If the state variable  $s_t = 1$  then Model 1 generates  $y_t$ , while if  $s_t = 2$  then Model 2 generates  $y_t$ . Both models are Markov processes with the following transition matrices:

Model 1	$y_{t+1}=UP$	$y_{t+1}=DOWN$	Model 2	$y_{t+1}=UP$	$y_{t+1}=DOWN$
$y_t=UP$	$\pi_L$	$1 - \pi_L$	$y_t=UP$	$\pi_H$	$1 - \pi_H$
$y_t=DOWN$	$1 - \pi_L$	$\pi_L$	$y_t=DOWN$	$1 - \pi_H$	$\pi_H$

Here  $0 < \pi_L < 0.5 < \pi_H < 1$  and thus Model 1 is the reversal model while Model 2 is the continuation model. The state variable also follows a Markov process with transition matrix

	$s_{t+1} = 1$	$s_{t+1} = 2$
$s_t = 1$	$1 - \lambda_1$	$\lambda_1$
$s_t = 2$	$\lambda_2$	$1 - \lambda_2$

Using the above information and the observed past outcomes, the decision maker uses the following recursive formula for her posterior belief,  $q_t$  about being in the reversal regime.

Given  $q_{t-1} = Prob(s_{t-1} = 1 | y_{t-1}, y_{t-2}, \dots, y_0) = Prob(s_{t-1} = 1 | y_{t-1}, y_{t-2}, q_{t-2})$ ,  $q_t$  is computed as:

$$\frac{((1 - \lambda_1)q_{t-1} + \lambda_2(1 - q_{t-1}))Pr(y_t | s_t = 1, y_{t-1})}{((1 - \lambda_1)q_{t-1} + \lambda_2(1 - q_{t-1}))Pr(y_t | s_t = 1, y_{t-1}) + (\lambda_1 q_{t-1} + (1 - \lambda_2)(1 - q_{t-1}))Pr(y_t | s_t = 2, y_{t-1})}.^{24}$$

As proven in BSV, if  $y_{t-1} = (\neq)y_t$  then  $q_{t-1} < (>)q_t$ , i.e., the decision maker puts more weight on Model 2 when she observes consecutive like outcomes. In our experiment subjects observe eight outcomes and are asked to report the probability of the next, ninth, outcome being UP.

<sup>24</sup>For a derivation of the formula the reader should refer to BSV. The authors use  $q_0 = 0.50$ .

Given  $y_1^s, \dots, y_8^s$ , and the underlying parameters  $\pi_L, \pi_H, \lambda_1, \lambda_2$ , the probability that the decision makers assigns to  $y_9$  being UP is

$$Pr(y_9^s = UP | y_1^s, \dots, y_8^s, \pi_{L,i}, \pi_{H,i}, \lambda_{1,i}, \lambda_{2,i}) = Pr(y_9^s = UP | s_9 = 1)q_8 + Pr(y_9^s = UP | s_8 = 2)(1 - q_8),$$

where

$$Pr(y_9^s = UP | s_t = 1) = I_{\{y_8^s = UP\}} * ((1 - \lambda_1)\pi_L + \lambda_1\pi_H) + I_{\{y_8^s = DOWN\}} * ((1 - \lambda_1)(1 - \pi_L) + \lambda_1(1 - \pi_H)),$$

$$Pr(y_9^s = UP | s_t = 2) = I_{\{y_8^s = UP\}} * ((1 - \lambda_2)\pi_H + \lambda_2\pi_L) + I_{\{y_8^s = DOWN\}} * ((1 - \lambda_2)(1 - \pi_H) + \lambda_2(1 - \pi_L)).$$

$I_{\{y_8^s = y\}} = 1$  if  $y_8^s = y$ ,  $y \in \{UP, DOWN\}$ , and 0 otherwise.

BSV take  $q_1 = 0.5$ . We estimate  $q_1$  for each subject.

## B. Rabin Model

Here we present the formal treatment of the Rabin model as is applied in our empirical analysis. Let the urn size be  $N$  and the renewal rate be  $K$ ,  $K < N$ . The possible values of the urn rates are taken to be

$$\Theta = \{\theta_1, \theta_2, \dots, \theta_9\} = \{10\%, 20\%, 30\%, 40\%, 50\%, 60\%, 70\%, 80\%, 90\%\}$$

Let  $\pi_j$  denote the prior probability  $Pr(\theta_j)$ ,  $j = 1, \dots, 9$ . Given a sequence  $y_1, y_2, \dots, y_n$  of  $n$  observed outcomes, let  $I$  denote the number UP outcomes in that sequence. Assuming that the urn has not been renewed (i.e., assuming that  $K > n$ ), the posterior probabilities of the possible urn rates are

$$\hat{\pi}_j = Pr(\theta_j | I) = \frac{Pr(I | \theta_j)\pi_j}{Pr(I)}, \quad (\text{A-1})$$

$$\text{where } Pr(I) = \sum_{j=1}^9 Pr(I | \theta_j)\pi_j, \quad \text{and } Pr(I | \theta_j) = \begin{cases} \frac{\binom{\theta_j N}{I} \binom{N - \theta_j N}{n - I}}{\binom{N}{n}} & \text{if } I \leq \theta_j N < N + I - n, \\ 0 & \text{otherwise.} \end{cases}$$

$$Pr(y_{n+1} = UP|y_1, y_2, \dots, y_n) = Pr(y_{n+1} = UP|I) = \sum_{j=1}^9 Pr(y_{n+1} = UP|I, \theta_j) Pr(\theta_j|I),$$

where  $Pr(y_{n+1} = UP|I, \theta_j) = \max(0, \frac{\theta_j^{N-I}}{N-n})$ .

Subjects in the experiment observe sequences of eight outcomes  $(y_1^s, \dots, y_8^s)$ . If  $K > 8$  then the above formula applied for  $n = 8$  provides  $Pr(y_9^s = UP|y_1^s, \dots, y_8^s, N_i, K_i, q_i, r_i)$ . If  $K \leq 8$  then the formula (A-1) has to be applied  $\lfloor \frac{8}{K} \rfloor^{25}$  times for  $n = K$  and one time for  $n = \text{mod}(8, K)$ , where with each application the posterior from the previous application becomes the prior of the next one. For example, if  $K=3$  then  $\lfloor \frac{8}{K} \rfloor = 2$  and  $\text{mod}(8, K) = 2$ . Thus, by first applying (A-1) with a prior  $\pi_j$ ,  $j = 1, \dots, 9$ , and the first three outcomes of the sequence (i.e,  $n = K = 3$ ), one derives the posterior  $\hat{\pi}_j$ ,  $j = 1, \dots, 9$ . Then using  $\hat{\pi}$  as the prior (for the second application of formula (A-1)), and outcomes four to six in the sequence (i.e,  $n = K = 3$ ), one derives the posterior  $\hat{\hat{\pi}}_j$ ,  $j = 1, \dots, 9$ . Finally, using  $\hat{\hat{\pi}}_j$  as the prior, one applies the above formula for the seventh and eighth outcomes (i.e.,  $n = \text{mod}(8, 3) = 2$ ) to obtain  $Pr(y_9^s = UP|y_1^s, \dots, y_8^s, N_i, K_i, q_i, r_i) = Pr(y_9^s = UP|y_7^s, y_8^s, N_i, \hat{\hat{\pi}}_j)$ .

---

<sup>25</sup> $\lfloor \cdot \rfloor$  denotes the floor function,  $\lfloor x \rfloor = \max\{a \in \mathbb{Z}, a \leq x\}$ .

## Tables and Figures

**Table I**  
**Basic Descriptive Statistics<sup>a</sup>**

		Reported Prob(UP)	Reaction Measure
Treatment 1	<i>Median</i>	0.50	0
	<i>Mean</i>	0.519	0.006
	<i>Q1</i>	0.40	-0.10
	<i>Q3</i>	0.65	0.13
Treatment 2	<i>Median</i>	0.50	0
	<i>Mean</i>	0.52	0.006
	<i>Q1</i>	0.40	-0.14
	<i>Q3</i>	0.70	0.15

---

<sup>a</sup>The statistics presented in the table are computed across subjects and across observed sequences.

**Table II**  
**Structural Estimation Results, Treatment 1**

Panel A: BSV

Model	Statistic	$\pi_L^a$	$\pi_H^b$	$\lambda_1^c$	$\lambda_2^d$	RMSE
BSV	<i>Mean</i>	0.44	0.71	0.1	0.24	0.1760
	<i>Quartile 1</i>	0.43	0.50	0	0.02	0.1299
	<i>Median</i>	0.47	0.59	0.01	0.24	0.1934
	<i>Quartile 3</i>	0.50	0.98	0.11	0.5	0.2291

Panel B: Rabin

		N	K	q	r	RMSE
Rabin	<i>Mean</i>	n.a. <sup>e</sup>	7.63	0.017	0.031	0.1639
	<i>Quartile 1</i>	12.50	7	0	0	0.1093
	<i>Median</i>	20	9	0.003	0.018	0.1787
	<i>Quartile 3</i>	39.26	9	0.02	0.05	0.2167

Panel C: Benchmarks

		q	r	RMSE
Bayesian <sup>f</sup>	<i>Mean</i>	0.017	0.028	0.1791
	<i>Quartile 1</i>	0	0	0.1418
	<i>Median</i>	0	0.0018	0.1924
	<i>Quartile 3</i>	0.03	0.04	0.2328
Correct model (Random Walk)	<i>Mean</i>			0.1906
	<i>Quartile 1</i>			0.1418
	<i>Median</i>			0.2077
	<i>Quartile 3</i>			0.2565

<sup>a</sup> $\pi_L$  is the probability of continuation in the reversal regime,  $0 \leq \pi_L \leq 0.5$ .

<sup>b</sup> $\pi_H$  is the probability of continuation in the continuation regime,  $0.5 \leq \pi_H \leq 1$ .

<sup>c</sup> $0 \leq \lambda_1 \leq 1$

<sup>d</sup> $0 \leq \lambda_2 \leq 1$

<sup>e</sup>Five of the subjects in Treatment 1 reported 50% after each sequence. Consequently, the upper bound on N is binding for those subjects (as their behavior is obtained as  $N \rightarrow \infty$ ). Even after the removal of those subjects, there are still others for whom the optimization procedure ends having reached the upper bound for  $N=10,000$ .

<sup>f</sup>The Bayesian model can be derived as a particular case of the Rabin model when  $N \rightarrow \infty$  or when  $K \rightarrow 1$ .

**Table III**  
**Simulation Results: Signed Reaction Averages<sup>a</sup>**

Streak Length								
1	2	3	4	5	6	7	8	
Panel A: Experimental Subjects								
0.7514	-0.1214	-0.5942	-0.6304	3.2880	2.5924	9.2174	8.9348	( $\times 10^{-2}$ )
Panel B: Rabin Simulations <sup>b</sup>								
-0.4256	-1.3605	-1.4070	-0.0992	1.0313	1.6549	7.9365	14.0945	( $\times 10^{-2}$ )
Panel C: BSV Simulation <sup>c</sup>								
-2.0160	0.1273	2.0977	3.5310	6.2593	8.4092	11.1173	15.9177	( $\times 10^{-2}$ )
Panel D: Bayesian Simulation <sup>d</sup>								
-0.0426	0.6129	2.3126	4.3253	5.3682	6.7173	9.5255	15.3538	( $\times 10^{-2}$ )

---

<sup>a</sup>Given the signed reactions of the experimental subjects and the signed reactions simulated according to BSV, the Rabin, and the Bayesian model, this table presents the averaged signed reactions grouped by sequence-ending streak length. Thus, given the sequences  $s_1, \dots, s_{100}$ , and letting  $|s_j|$  denote the length of the streak ending the sequence  $s_j$ ,  $j = 1, \dots, 100$ , the table reports  $\frac{1}{L} \sum_{i,j, |s_j|=k} Y_{ij}$ ,  $k = 1, \dots, 8$ , where  $Y_{ij}$  is the reaction measure of subject  $i$  for sequence  $s_j$ .  $L$  is the number of summands.

<sup>b</sup>The panel reports the aggregated results from the subject-level Rabin simulations.

<sup>c</sup>The panel reports the aggregated results from the subject-level BSV simulations.

<sup>d</sup>The panel reports the aggregated results from the subject-level Bayesian simulations.

**Table IV**  
**Piecewise Linear Regression, Treatment 1<sup>a</sup>**

		Coefficient Estimates <sup>b</sup>			
Cutoff	Variable	Data	Rabin	BSV	Bayesian
C=3	STREAK1	-0.912 (-2.148)	-0.794 (-4.322)	1.927 (18.759)	1.017 (8.260)
	STREAK2	1.742 (4.505)	2.032 (12.140)	2.327 (24.858)	1.992 (17.760)
	Adj $R^2$	0.0637 (10.19)	0.1431 (76.82)	0.6021 (1020.66)	0.1869 (375.80)
C=4	STREAK1	-0.509 (-1.561)	-0.362 (-2.564)	1.905 (24.159)	1.186 (12.541)
	STREAK2	2.534 (4.493)	2.869 (11.755)	2.614 (19.164)	2.240 (13.688)
	Adj $R^2$	0.0639 (10.63)	0.1440 (79.27)	0.6029 (1026.73)	0.1866 (375.07)

<sup>a</sup>The estimated (fixed effects) model is  $Y_{ij} = \mu_i + \beta_1 STREAK1_{ij} + \beta_2 STREAK2_{ij} + \epsilon_{ij}$ ,  $Y_{ij}$  is the reaction measure of subject  $i$  in trial  $j$ ,  $i = 1, \dots, 46$ ,  $j = 1, \dots, 100$ . We report the results for the experimental data, and for each of the simulated data sets corresponding to the Rabin, BSV, and the Bayesian model. We report results for  $C=2,3,4$ .

<sup>b</sup>The t-statistics and the F-statistics are in the parentheses.

**Table V**  
**Structural Estimation Results for the Rabin Model, Treatment 2**

		N	K	q	r	RMSE
Rabin	<i>Mean</i>	n.a.	7.09	0.01	0.021	0.1968
	<i>Quartile 1</i>	14.5	5	0	0	0.951
	<i>Median</i>	20	9	0.0003	0.004	0.2106
	<i>Quartile 3</i>	168.84	9	0.0025	0.03	0.2882
Bayesian	<i>Mean</i>			0.009	0.017	0.211
	<i>Quartile 1</i>			0	0	0.1055
	<i>Median</i>			0	0.001	0.2292
	<i>Quartile 3</i>			0.006	0.027	0.2951
Correct model (Random Walk)	<i>Mean</i>					0.2148
	<i>Quartile 1</i>					0.1058
	<i>Median</i>					0.2314
	<i>Quartile 3</i>					0.3030

**Table VI**  
**Signed Reaction Averages: Comparison between Treatments<sup>a</sup>**

Data	Streak Length								
	1	2	3	4	5	6	7	8	
Treatment 1	0.7514	-0.1214	-0.5942	-0.6304	3.2880	2.5924	9.2174	8.9348	( $\times 10^{-2}$ )
Treatment 2	1.7957	2.1667	-2.2373	-4.6522	-2.4239	-2.9348	-0.5652	-1.1304	( $\times 10^{-2}$ )

---

<sup>a</sup>Given the signed reactions of the experimental subjects in Treatment 1 and Treatment 2, this table presents the averaged signed reactions grouped by sequence-ending streak length. Thus, given the sequences  $s_1, \dots, s_{100}$ , and letting  $|s_j|$  denote the length of the streak ending the sequence  $s_j$ ,  $j = 1, \dots, 100$ , the table reports  $\frac{1}{L} \sum_{i,j,|s_j|=k} Y_{ij}$ ,  $k = 1, \dots, 8$ , where  $Y_{ij}$  is the reaction measure of subject  $i$  for sequence  $s_j$ .  $L$  is the number of summands. Using the eight averages, the paired t-test (the Wilcoxon Signed Rank Test) rejects the equality of the means (medians) of the samples in favor of the alternative that the subjects in Treatment 2 tend to predict more reversals, one-tailed p-value is 0.018 (0.027).

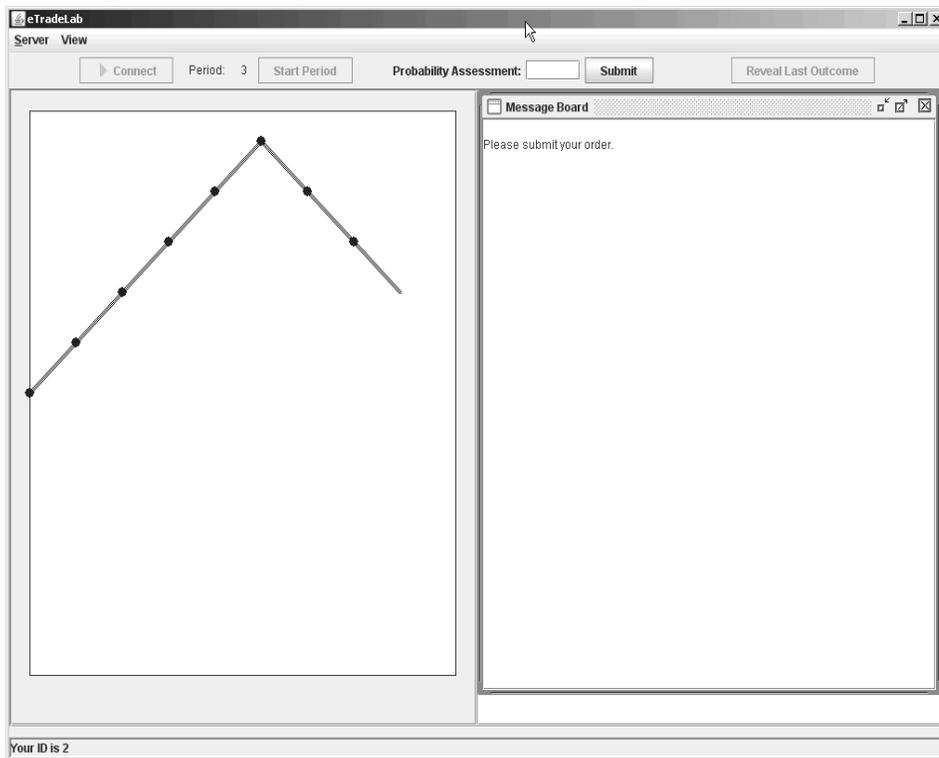
**Table VII**  
**Piecewise Linear Regression, Treatment 2<sup>a</sup>**

Coefficient Estimates <sup>b</sup>			
Cutoff	Variable	Data	Rabin
C=4	<i>STREAK1</i>	-1.885 (-4.949)	-1.332 (-10.082)
	<i>STREAK2</i>	0.537 (0.815)	1.501 (6.569)
	Adj $R^2$	0.0565 (14.71)	0.0809 (29.63)
C=5	<i>STREAK1</i>	-1.583 (-5.070)	-1.056 (-9.757)
	<i>STREAK2</i>	1.247 (1.220)	2.698 (7.610)
	Adj $R^2$	0.0562 (14.00)	0.0806 (28.96)

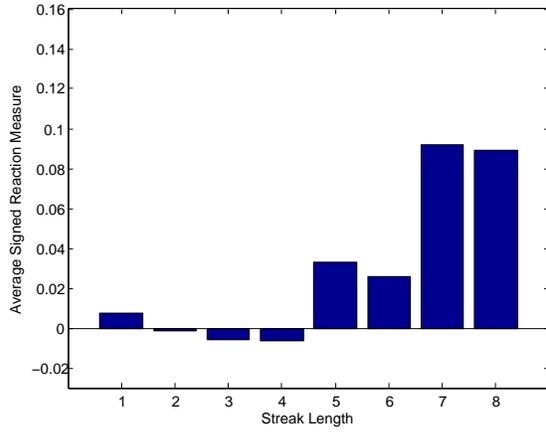
<sup>a</sup>The estimated (fixed effects) model is  $Y_{ij} = \mu_i + \beta_1 STREAK1_{ij} + \beta_2 STREAK2_{ij} + \epsilon_{ij}$ ,  $Y_{ij}$  is the reaction measure of subject  $i$  in trial  $j$ ,  $i = 1, \dots, 46$ ,  $j = 1, \dots, 100$ . We report the results for the experimental data and for the simulated data sets corresponding to the Rabin model. We report results for  $C=4$  and 5.

<sup>b</sup>The t-statistics and the F-statistics are in the parentheses.

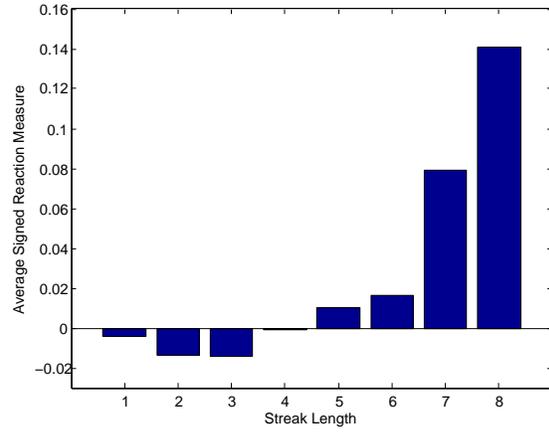
Figure 1. A Snapshot of the Interactive Experimental Software



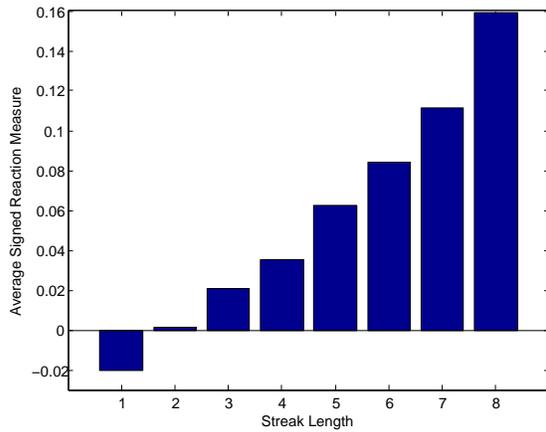
**Figure 2.** Reaction Measure-Streak Length Graphs <sup>a</sup>



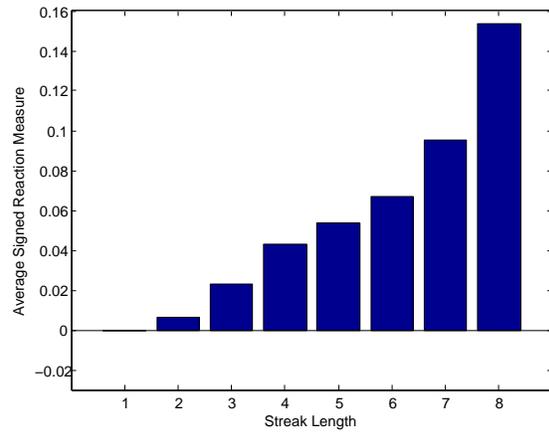
(a) The Experimental Data



(b) The Rabin Model



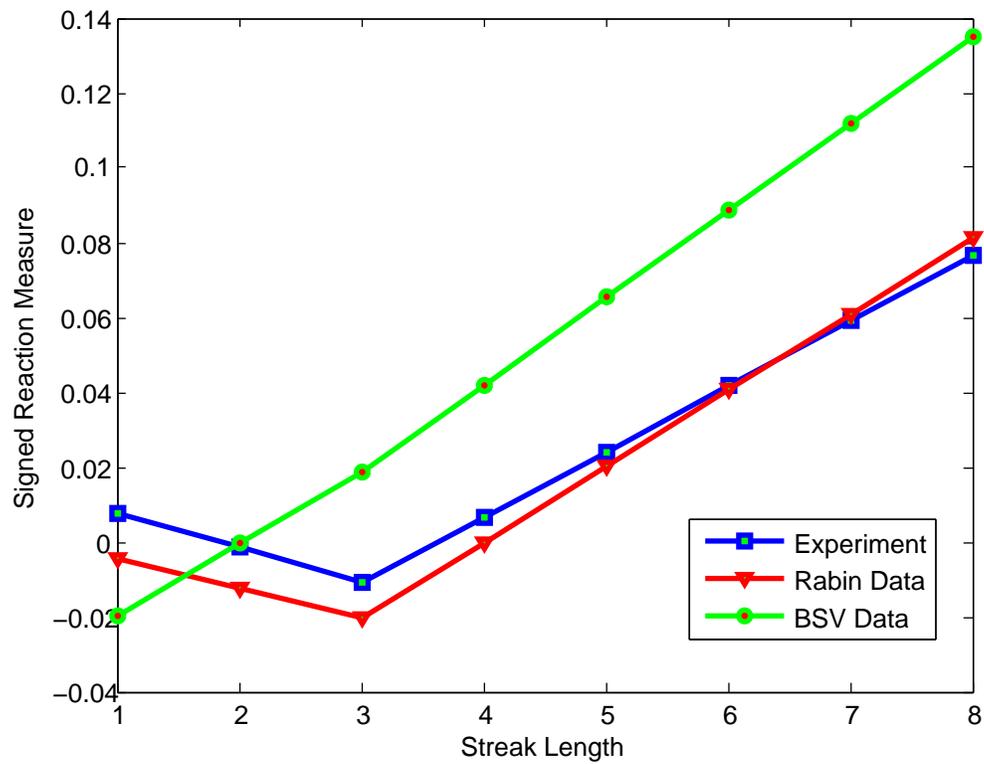
(c) The BSV Model



(d) The Bayesian Model

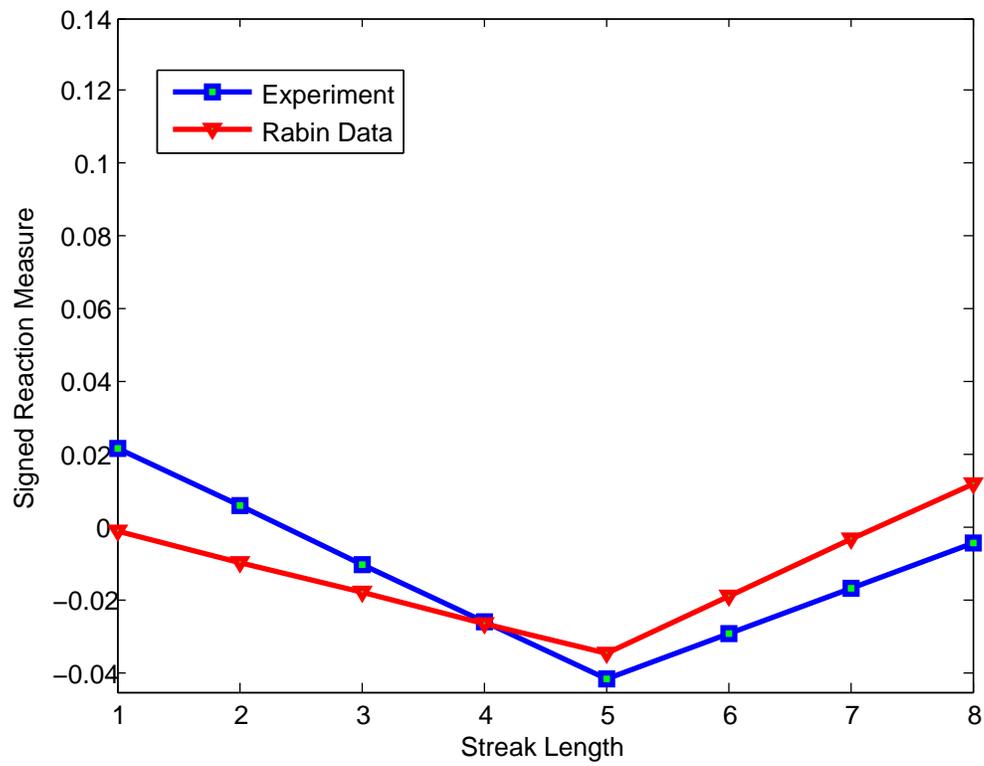
<sup>a</sup>Ending streak length is on the x-axis. Average signed reaction measure (averaged across all subject's responses for the corresponding streak length) is on the y-axis.

Figure 3. Treatment 1 Regression Results<sup>a</sup>



<sup>a</sup>The figure graphically displays the results of the regression  $Y_{ij} = \mu_i + \beta_1 STREAK1_{ij} + \beta_2 STREAK2_{ij} + \epsilon_{ij}$  (where  $Y_{ij}$  is the reaction measure of subject  $i$  in trial  $j$ ,  $i = 1, \dots, 46$ ,  $j = 1, \dots, 100$ ) as presented in Table IV. The graph is for a cutoff of  $C=3$ .

Figure 4. Treatment 2 Regression Results<sup>a</sup>



<sup>a</sup>The figure graphically displays the results of the regression  $Y_{ij} = \mu_i + \beta_1 STREAK1_{ij} + \beta_2 STREAK2_{ij} + \epsilon_{ij}$  (where  $Y_{ij}$  is the reaction measure of subject  $i$  in trial  $j$ ,  $i = 1, \dots, 46$ ,  $j = 1, \dots, 100$ ) as presented in Table VII. The graph is for a cutoff of  $C=4$ .

## References

- Barberis, Nicholas, Andrei Shleifer, and Robert Vishny, 1998, A Model of Investor Sentiment, *Journal of Financial Economics* 49, 307–43.
- Bar-Hillel, Maya, and Willem A. Wagenaar, 1991, Perception of Randomness, *Advances of Applied Mathematics* XII, 428–54.
- Bloomfield, Robert, and Jeffrey Hales, 2002, Predicting the Next Step of a Random Walk: Experimental Evidence of Regime-Shifting Beliefs, *Journal of Financial Economics* 65, 397–414.
- Brav, Alon and John B. Heaton, 2002, Competing Theories of Financial Anomalies, *Review of Financial Studies*, 11, 179-184.
- Bruine de Bruin, Wändi, Paul S. Fischbeck, Neil A. Stilber, and Baruch Fischhoff, 2002, What Number is “Fifty-Fifty”? : Redistributing Excessive 50% Responses in Elicited Probabilities *Risk Analysis*, 22, 713-723.
- Burns, Bruce D. and Bryan Corpus, 2004, Randomness and Induction from Streaks: “Gambler’s Fallacy” versus “Hot Hand,” *Psychonomic Bulletin and Review*, 11, 179-184.
- Clotfelter, Charles, and Philip Cook, 1993, The ‘Gambler’s Fallacy’ in Lottery Play, *Management Science* 39, 1521–25.
- Croson, Rachel, and Jim Sundali, 2005, The Gambler’s Fallacy and the Hot Hand: Empirical Data from Casinos, *Journal of Risk and Uncertainty* 30, 195–209.
- Durham, Greg, Michael Hertzel and Spencer Martin, 2005, The Market Impact of Trends and Sequences in Performance: New Evidence, *Journal of Finance* 60, 2551–2569.
- Fox, Craig R. and Yuval Rottenstreich, 2003, Partition Priming in Judgement Under Uncertainty, *Psychological Science* 14, 195–200.
- Massey, Cade, and George Wu, 2005, Detecting Regime Shifts: The Causes of Under- and Overreaction, *Management Science* 51, 932–947.
- Mullainathan, Sendhil, 2002, Thinking Through Categories, working paper.

- Offerman, Theo, and Joep Sonnemans, 2001, Is the Quadratic Scoring Rule really incentive compatible?, working paper, <http://dare.uva.nl/document/150>.
- Offerman, Theo, and Joep Sonnemans, 2004, What's Causing Overreaction? An Experimental Investigation of Recency and the Hot Hand Effect, *Scandinavian Journal of Economics* 106, 533-553.
- Rabin, Matthew, 2002, Inference by Believers in the Law of Small Numbers, *The Quarterly Journal of Economics* 117, 775-816.
- Rabin, Matthew and Dimitri Vayanos, January 2007, The Gambler's and Hot-Hand Fallacies: Theory and Applications, <http://ssrn.com/abstract=954636>.
- Rapoport, Amnon, and David V. Budescu, 1992, Generation of Random Series in Two-Person Strictly Competitive Games, *Journal of Experimental Psychology: General* CXXI, 352-363.
- Terrell, Dek, 1994, A Test of Gambler's Fallacy—Evidence from Para-Mutuel Games *Journal of Risk and Uncertainty* VIII, 603-617.
- Tversky, Amos and Daniel Kahneman 1971, Belief in the Law of Small Numbers, *Psychological Bulletin* LXXVI, 105-110.