

# On the Robustness and Provenance of the Gambler's Fallacy

Psychological Science  
2025, Vol. 36(6) 451–464  
© The Author(s) 2025  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/09567976251344570  
www.psychologicalscience.org/PS



Yang Xiang<sup>1</sup>, Kevin Dorst<sup>2</sup>, and Samuel J. Gershman<sup>1,3,4</sup>

<sup>1</sup>Department of Psychology, Harvard University; <sup>2</sup>Department of Linguistics and Philosophy, Massachusetts Institute of Technology; <sup>3</sup>Center for Brain Science, Harvard University; and <sup>4</sup>Center for Brains, Minds, and Machines, Massachusetts Institute of Technology

## Abstract

The gambler's fallacy is typically defined as the false belief that a random event is less likely to occur if it has occurred recently. Although forms of this fallacy have been documented numerous times, past work either has not actually measured probabilistic predictions but rather point predictions or used sequences that were not independent. To address these problems, we conducted a series of high-powered, preregistered studies in which we asked 750 adult Amazon Mechanical Turk workers from the United States to report probabilistic predictions for truly independent sequences. In contrast to point predictions, which generated a significant gambler's fallacy, probabilistic predictions were not found to lead to a gambler's fallacy. Moreover, the point predictions could not be reconstructed by sampling from the probability judgments. This suggests that the gambler's fallacy originates at the decision stage rather than in probabilistic reasoning, as posited by several leading theories. New theories of the gambler's fallacy may be needed to explain these findings.

## Keywords

gambler's fallacy, probability judgment, probabilistic reasoning, point prediction, sampling, decision heuristic

Received 10/19/24; Revision accepted 5/3/25

The gambler's fallacy is a false belief that a random event is less likely to occur if the event has occurred recently, even when the probability of the event is objectively known to be independent from trial to trial (Clotfelter & Cook, 1993; Suetens & Tyran, 2012). For example, when observing a sequence of coin tosses, people expect tails to be more likely after a run of heads. Ever since the fallacy was first proposed by Marquis de Laplace (1902), much work has documented it in real-world contexts, including casino betting (Sundali & Croson, 2006), lottery play (Clotfelter & Cook, 1993; Kong et al., 2020; Suetens & Tyran, 2012; Terrell, 1994), penalty shootouts (Misirlisoy & Haggard, 2014), and even predictions of next-child gender (McClelland & Hackenberg, 1978), as well as in many laboratory experiments (Ayton & Fischer, 2004; Barron & Leider, 2010; Rao & Hastie, 2023; Roney & Sansone, 2015; Roney & Trick, 2003; Tversky & Kahneman, 1974).<sup>1</sup>

Researchers have proposed a myriad of theories to explain this fallacy. The existing theories largely fall under two classes. One class of models assumes it arises from irrational systematic errors that result from a reliance on the representativeness heuristic and the law of small numbers (Kahneman & Tversky, 1972; Miller & Sanjurjo, 2018; Tversky & Kahneman, 1971, 1974). According to these accounts, people regard a sample randomly drawn from a population as highly representative and similar to the population in all essential characteristics. This belief in local representativeness creates the gambler's fallacy because reversals (e.g., the occurrence of black after observing a long run of red on a roulette wheel) will restore the balance and result in a

## Corresponding Author:

Yang Xiang, Department of Psychology, Harvard University  
Email: yyx@g.harvard.edu

more representative sequence than repetition (e.g., the occurrence of another red). The other class of models posits that the fallacy arises from rational probabilistic reasoning about the underlying generator. Theories within this second class can further be grouped into two types: *Error-optimal models* assume that people rationally act on their mistaken model of the world, such as assuming that the generator changes over time (Barberis et al., 1998; Rabin, 2002; Rabin & Vayanos, 2010), whereas *bounded-optimal models* assume that the fallacy occurs because people have limited memory (Dorst, 2024; Farmer et al., 2017; Hahn & Warren, 2009).

Despite the differences in the specific formalism and the assumptions, all of these theories assume that the fallacy stems from subjective probability judgments that then get translated into point predictions through a decision function—such as guessing the outcome that is most likely. Here, by “point predictions,” we mean a single-outcome prediction (e.g., heads or tails in the case of a coin), consistent with the literature (Agliari et al., 2016; Glazer & Rubinstein, 2024; Tergiman, 2015; Weitzel & Rosenkranz, 2016; Weiss & Shanteau, 2004). In other words, these theories all postulate that (a) subjective probability judgments should exhibit the gambler’s fallacy and (b) probability judgments and point predictions should move together.

Remarkably, almost none of the past work demonstrating the fallacy—whether in the real world or laboratory experiments—has tested these postulates directly. For a study to truly be a test of the gambler’s fallacy, it has to measure probability judgments based on independent events. Past work based on real-world data has examined only point predictions (Clotfelter & Cook, 1993; Kong et al., 2020; McClelland & Hackenberg, 1978; Misirlisoy & Haggard, 2014; Suetens & Tyran, 2012; Sundali & Croson, 2006; Terrell, 1994), as is the case with most of the laboratory studies (Ayton & Fischer, 2004; Barron & Leider, 2010; Jones, 1971; Roney & Sansone, 2015; Roney & Trick, 2003; Tversky & Kahneman, 1974). Some laboratory studies have elicited probability judgments, but they used nonindependent sequences (Bloomfield & Hales, 2002; Rao & Hastie, 2023). One exception is Asparouhova et al. (2009), who showed participants eight-outcome sequences generated independently from a Bernoulli distribution and asked for probability judgments. However, a closer look at participants’ reported probability that a streak would repeat ( $N = 46$ ; 100 trials each) revealed that the median was .50 and the mean was .506, indicating that there was barely any gambler’s fallacy. Thus, evidence for a true gambler’s fallacy in probabilistic beliefs is weak to nonexistent. If indeed a true gambler’s fallacy does not exist in probabilistic beliefs, that would raise questions as to whether the existing theories rest on an erroneous

premise and call for new theories that do not rely on access to probabilistic reasoning and a new understanding of the mechanisms that drive the gambler’s fallacy.

In the current work, we tested the gambler’s fallacy systematically, adhering to its probabilistic definition: Adapting the materials from Rao and Hastie’s (2023) bingo-ball color-prediction task, we told participants the objective ground truth probability, showed participants independent sequences, and elicited probability judgments from them. Measuring probability judgments directly allowed us to investigate whether the gambler’s fallacy originates from probability judgments, as postulated by the existing theories. In Experiments 1a and 1b ( $N = 300$ ), we used independent and identically distributed (IID) sequences with 50% probability of producing each outcome (blue or red bingo ball). Experiments 2a and 2b ( $N = 300$ ) repeated these experiments using other probabilities (40% and 60%). Finally, Experiment 3 replicated an experiment from Rao and Hastie (2023) with nonindependent sequences. These studies collectively allowed us to interrogate the robustness and provenance of the gambler’s fallacy.

## Research Transparency Statement

### General disclosures

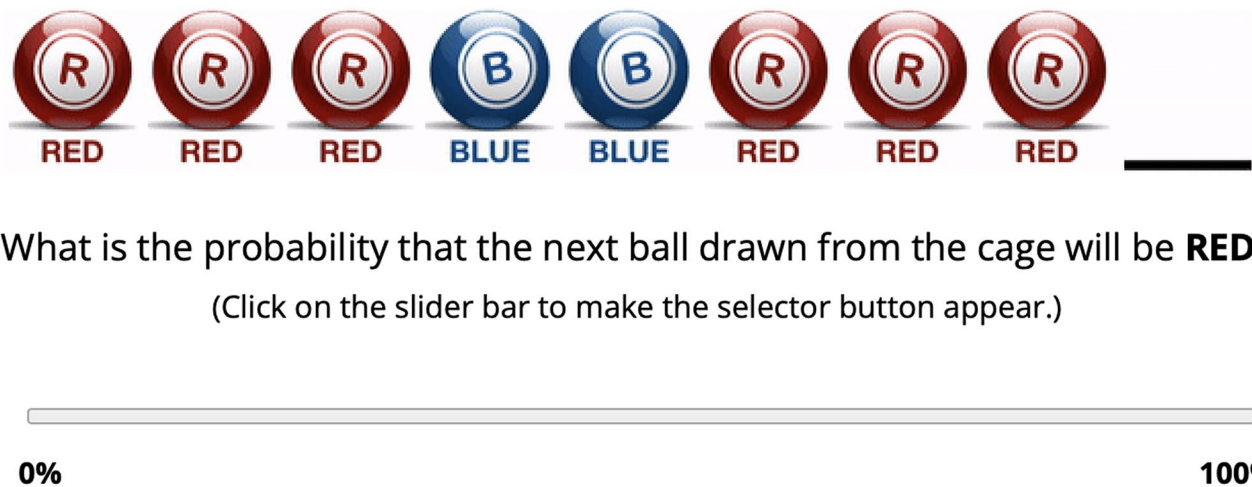
**Conflicts of interest:** All authors declare no conflicts of interest. **Funding:** This research was funded by National Science Foundation Grant DRL-2024462. **Artificial intelligence:** No AI-assisted technologies were used in this research or the creation of this article. **Ethics:** This research received approval from the Harvard Institutional Review Board (Protocol No. IRB15-2048). **Open Science Framework:** To ensure long-term preservation, all OSF files were registered at <https://doi.org/10.17605/OSF.IO/VRU6Y>.

### Experiments 1a and 1b disclosures

**Preregistration:** The research questions, method, and analyses were preregistered (<https://aspredicted.org/2frb-zshx.pdf>) on June 29, 2024, prior to data collection, which began on July 1, 2024. There were no deviations from the preregistration. **Materials:** All study materials are publicly available (<https://osf.io/67nsd>). **Data:** All primary data are publicly available (<https://osf.io/67nsd>). **Analysis scripts:** All analysis scripts are publicly available (<https://osf.io/67nsd>).

### Experiments 2a and 2b disclosures

**Preregistration:** The research questions, method, and analyses were preregistered (<https://aspredicted.org/>)



**Fig. 1.** Example trial in Experiment 1a. Participants observed eight balls drawn from a bingo machine one after another with replacement and predicted the color of the ninth ball.

prws-hqh3.pdf) on July 17, 2024, prior to data collection, which began on July 23, 2024. There were no deviations from the preregistration. **Materials:** All study materials are publicly available (<https://osf.io/67nsd>). **Data:** All primary data are publicly available (<https://osf.io/67nsd>). **Analysis scripts:** All analysis scripts are publicly available (<https://osf.io/67nsd>).

### Experiment 3 disclosures

**Preregistration:** The research questions, method, and analyses were preregistered (<https://aspredicted.org/khp6-hkz8.pdf>) on July 9, 2024, prior to data collection, which began later that same day. There were no deviations from the preregistration. **Materials:** All study materials are publicly available (<https://osf.io/67nsd>). **Data:** All primary data are publicly available (<https://osf.io/67nsd>). **Analysis scripts:** All analysis scripts are publicly available (<https://osf.io/67nsd>).

### Experiments 1a and 1b

In these experiments, we modified the task design of predicting colors of bingo balls from Experiments 2a and 2b of Rao and Hastie (2023), replacing the stimuli with IID sequences (50% probability of producing each outcome, a blue or a red ball). The two experiments differed in the type of response we asked participants to report: Experiment 1a elicited probability judgments, whereas Experiment 1b elicited point predictions.

We note here that much of the prior experimental work on the gambler's fallacy (with the notable exception of Rao and Hastie, 2023) is based on a few dozen

participants per experiment, with few replications or preregistrations. To ensure that our experiments are reliably able to detect even small effects, we used a large sample size (150 participants per experiment), and all of our experiments are preregistered replications of pilot studies.

### Method

**Participants.** We recruited 150 participants for each experiment via Amazon's Mechanical Turk (MTurk) platform. After reading the instructions, participants completed a comprehension check that tested their understanding of the task, with special attention to ensuring they understood that the balls were sampled with replacement. Participants received \$2.50 to complete eight trials. The experiments were approved by the Harvard Institutional Review Board and preregistered at <https://aspredicted.org/2frb-zshx.pdf>. For these and subsequent experiments, we did not exclude any participants or observations.

**Stimuli.** The stimuli were 18 IID sequences of eight random ball colors generated on the fly on each trial for each participant. Each ball color was independently sampled from a Bernoulli distribution with  $p = .5$  of being blue or red.

**Procedure.** Participants completed a total of 18 trials (for an example trial, see Fig. 1). On each trial, eight bingo balls were drawn, one after another with a 1-s delay, by a mechanical bingo machine from a covered cage with 50 blue balls and 50 red balls. Participants were told that each time after a ball is drawn the machine

would roll the ball back to the cage and spin the cage before the next ball is drawn. After seeing eight draws, participants predicted the color of the next ball. In Experiment 1a, participants reported probability judgments (the probability of the next ball being red) using a slider. In Experiment 1b, participants reported point prediction by choosing between red and blue.

## Results

Figure 2 shows the participants' responses to the eight most common sequences. Overall, for the same sequence, participants predicted repetition more for probability judgments and reversal more for point predictions.

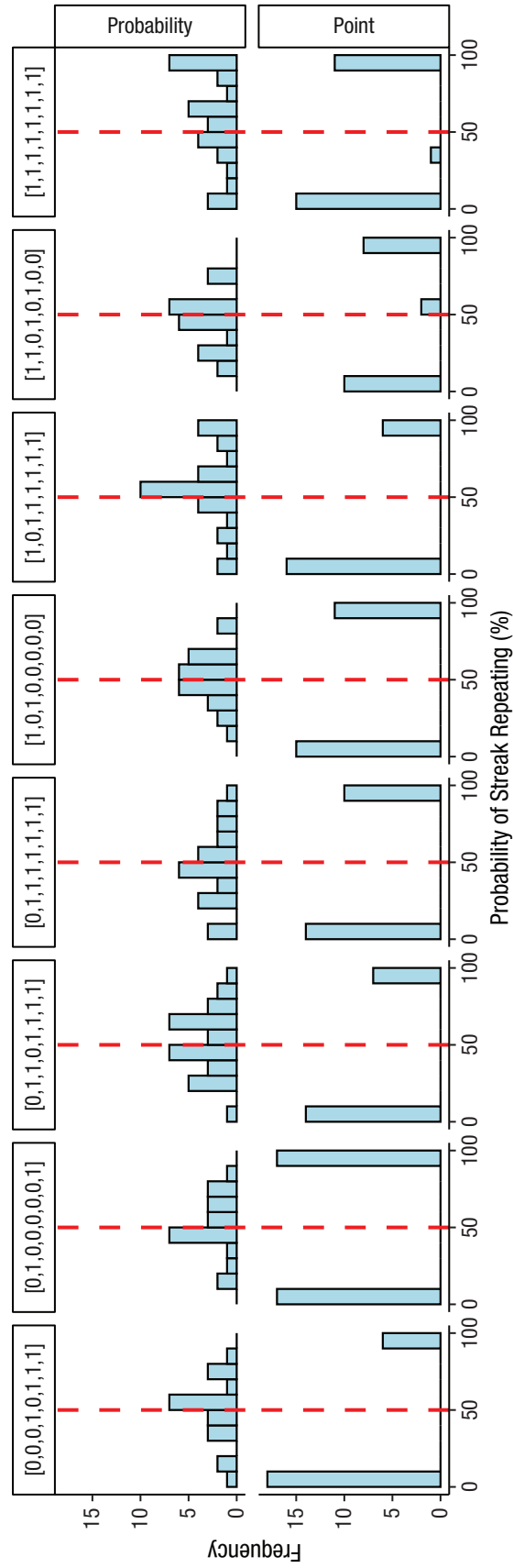
We computed the probability that a streak will repeat— $P(\text{repeat})$ —for each participant by averaging their predictions across trials. Point predictions from Experiment 1b were treated as 0% and 100% for this calculation to compute  $P(\text{repeat})$  as the proportion of trials in which they predicted that the streak would repeat. Note that the meaning of this average is conceptually different for the two experiments and we are not directly comparing the two. For the following analyses, we started with a normality test that determined whether we used a parametric  $t$  test or nonparametric Wilcoxon signed-rank test. We were originally concerned about normality when we reanalyzed Rao and Hastie's (2023) data (see Fig. 3), which showed a skewed distribution indicating that the effect was mainly driven by a small group of participants. A Lilliefors (Kolmogorov-Smirnov) test for normality indicated that the data were not normally distributed ( $D = 0.13$ ,  $p < .001$ ). This finding motivated us to check for normality first. Note that given our large sample sizes ( $N = 150$  for each experiment), we could apply the  $t$  test even if the data were not normally distributed. We additionally reported Bayesian  $t$  test results so that we could evaluate the evidence for the null hypothesis.

For Experiment 1a, in which participants reported probability judgments, a Lilliefors (Kolmogorov-Smirnov) test for normality indicated that the data were not normally distributed ( $D = 0.15$ ,  $p < .001$ ). Therefore, we conducted a one-sample Wilcoxon signed-rank test, which indicated that the median was not significantly different from 50% ( $Z = -1.03$ ,  $p = .304$ ,  $r = .08$ ). A Bayesian  $t$  test yielded a BF ( $\text{BF}_{10}$ ) of 0.248, meaning that the data were four times more likely under the null hypothesis that the mean would not be different from 50%. The median resulting posterior distribution for the effect size ( $\delta$ ) was  $-0.11$ —95% credible interval (CrI) =  $[-0.27, 0.05]$ . See the first row of Figure 4 (left) for the distribution of  $P(\text{repeat})$ . The first row of Figure 5a

shows the mean and median predictions by terminal streak length, as in Rao and Hastie (2023).

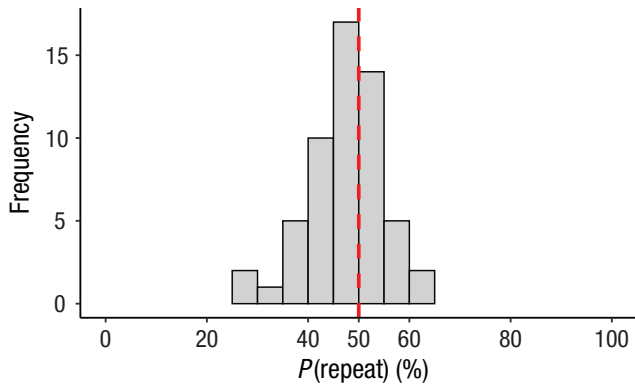
By contrast, we observed more probability mass falling below 50% in Experiment 1b (Fig. 4, first row, right panel), in which participants reported point predictions. A Lilliefors (Kolmogorov-Smirnov) test for normality indicated that the data were not normally distributed ( $D = 0.11$ ,  $p < .001$ ). A one-sample Wilcoxon signed-rank test indicated that the median was significantly different from 50% ( $Z = -5.56$ ,  $p < .001$ ,  $r = .45$ ). A Bayesian  $t$  test yielded a  $\text{BF}_{10} > 1,000$ , meaning that the data were more than 1,000 times more likely under the alternative hypothesis that the mean would be different from 50%. The median resulting posterior distribution for the effect size ( $\delta$ ) was  $-0.51$ —95% CrI =  $[-0.68, -0.34]$ . The first row of Figure 5b shows the mean predictions by terminal streak length.

Past work has suggested that people often make decisions from a few samples (Icard, 2016; Vul et al., 2014), which may lead to systematic deviations from the probability judgments (Sanborn, 2017; Sanborn & Chater, 2016). In light of this idea, we conducted an additional analysis to find out whether the point predictions could be explained by a transformation of the probability judgments. In other words, could we recover the point prediction patterns in Experiment 1b by sampling from the probability judgments in Experiment 1a or from the ground truth probability? We formalized these two hypotheses as the *subjective probability model* and the *objective probability model*. For each trial in Experiment 1a, the subjective probability model implements an optional stopping rule (Zhu et al., 2024) from which it draws samples sequentially from a Bernoulli distribution with a mean given by participants' reported probability until there are five more samples supporting one prediction over the other. The objective probability model generates predictions using the same process except that the samples are drawn from a Bernoulli distribution with a mean given by the ground truth probability (50%). The resulting  $P(\text{repeat})$  is shown in the first row of Figure 5b. Sampling could not recover the point prediction patterns. This suggests that the point prediction patterns are likely generated by a process different from probability judgments. We additionally explored different thresholds for terminating the sampling process, results of which are shown in Figure S1 in the Supplemental Material available online. These models still could not recover the point prediction patterns. We also considered a *thresholding model* that converts probability judgments to point predictions by applying a cutoff at the ground truth (which is effectively drawing an infinite number of samples, as illustrated in Fig. S1). The model could not capture the



**Fig. 2.** Average probability that a streak will repeat for the most common sequences in Experiments 1a and 1b. Each panel corresponds to an eight-ball sequence: 0 = blue ball, 1 = red ball. The red dashed lines mark the ground truth (50%).



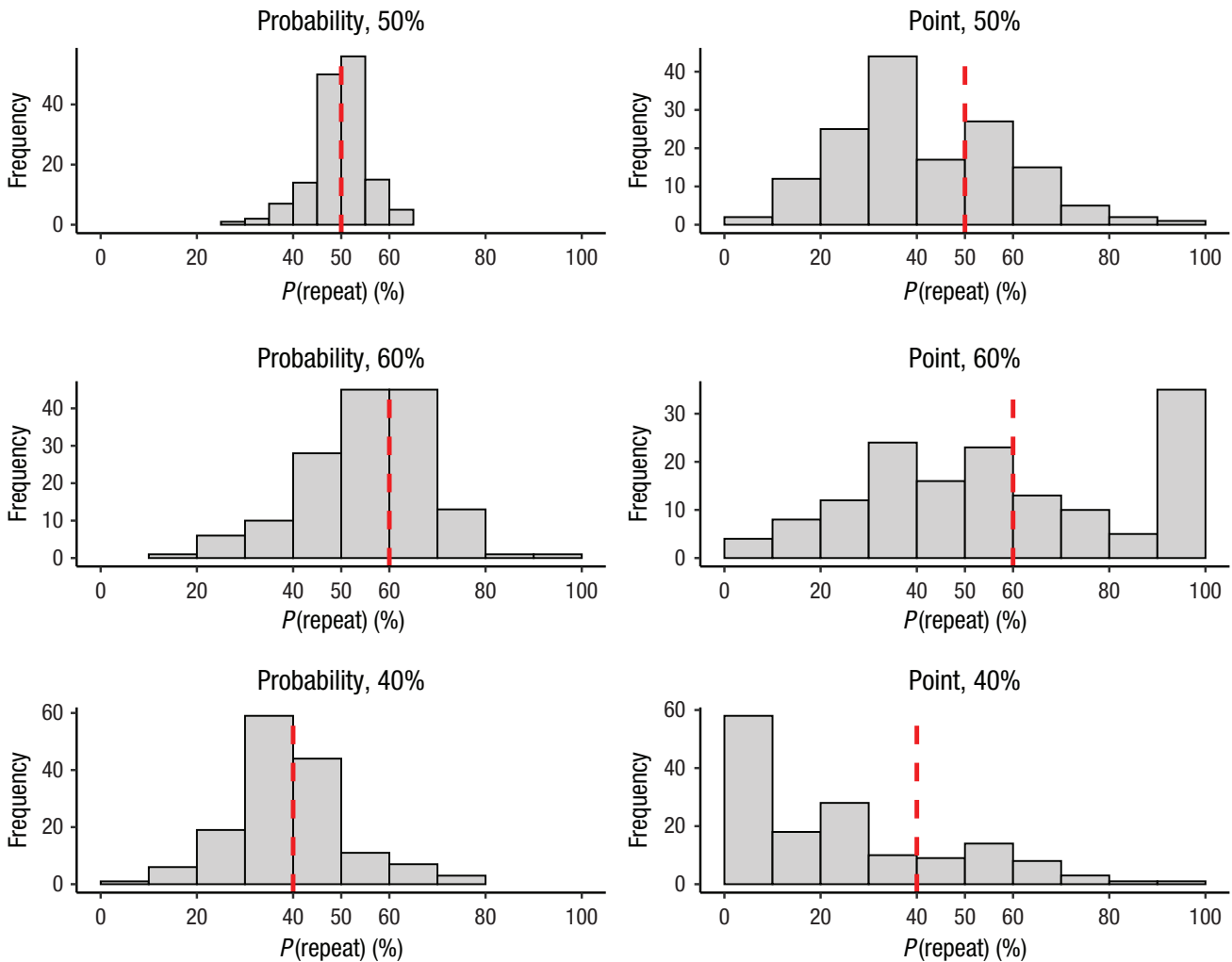


**Fig. 3.** Reanalysis of Rao and Hastie (2023) Study 2A, bingo50 condition, showing the average probability that a streak will repeat. The dashed red line marks the ground truth probability (50%).

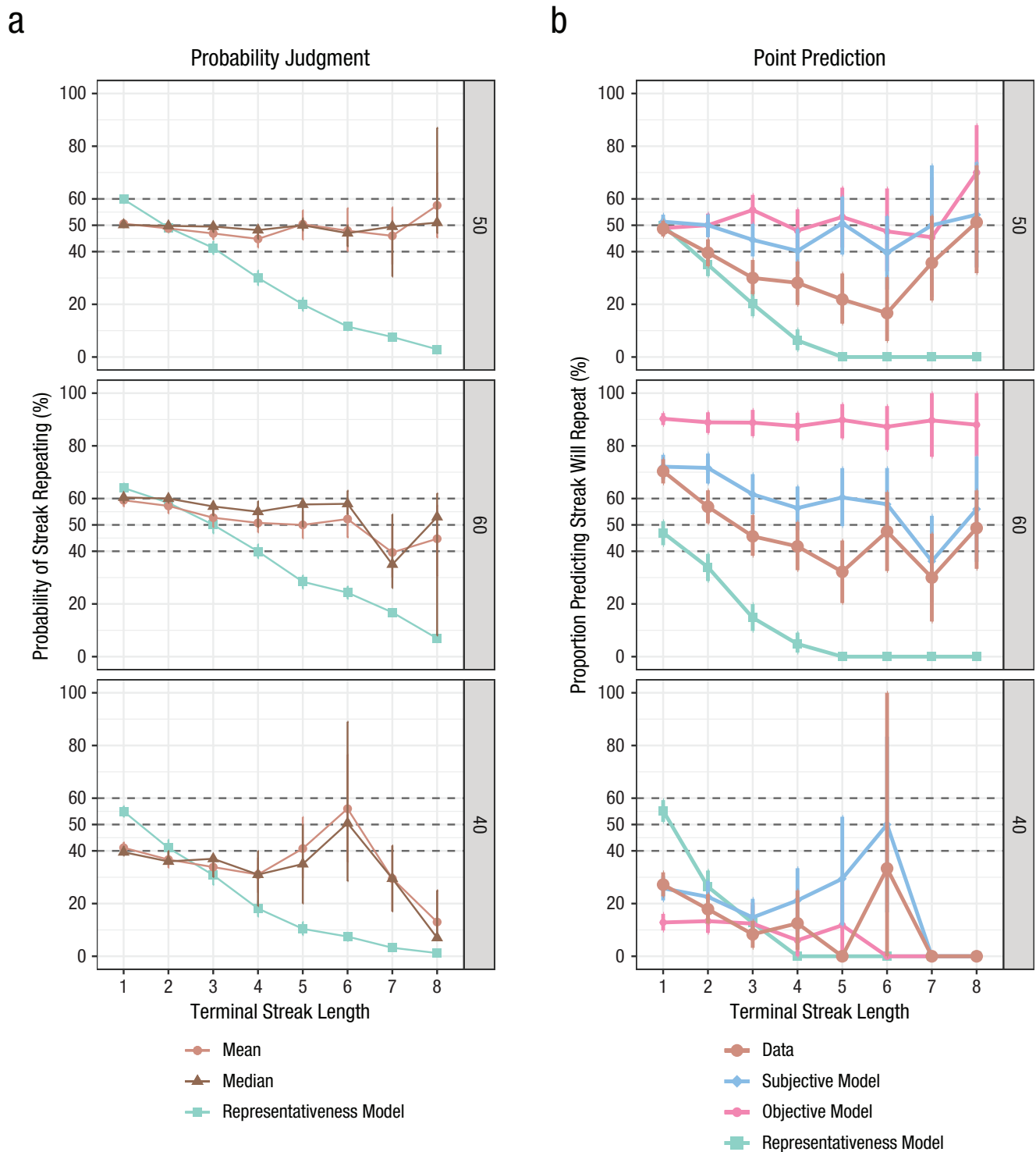
magnitude of participants' responses. We conducted a Bayesian  $t$  test comparing participants'  $P(\text{repeat})$  in

Experiment 1b and the thresholding model's point predictions on the basis of participants' probability judgments in Experiment 1a, which yielded a  $BF_{10}$  of 21.514, meaning that the data were more than 20 times more likely under the alternative hypothesis that the mean of participants'  $P(\text{repeat})$  would be different from the model's predicted  $P(\text{repeat})$ . The median resulting posterior distribution for the effect size ( $\delta$ ) was  $-0.37$ —95% CrI =  $[-0.59, -0.14]$ .

Could this disconnection be explained by the representativeness heuristic? To find out, we explored a *representativeness model*, which makes probability judgments and point predictions to render the sequence more representative in reflecting the ground truth probability. For example, after observing eight blue balls in a row, the representativeness model would predict that a red ball is “due” and thus that a red ball is more likely to appear in the next round than a blue ball. For probability judgments, we computed the number of balls



**Fig. 4.** Average probability that a streak will repeat. The title of each subplot indicates the type of response elicited (“probability” or “point”) and the ground truth probabilities (50% in Experiments 1a and 1b, 60% and 40% in Experiments 2a and 2b). The dashed red lines mark the ground truth probabilities.



**Fig. 5.** Probability that a streak will repeat for different terminal streak lengths. Mean and median responses and the representativeness model are shown for (a) Experiments 1a and 2a. The data, subjective probability model, objective probability model, and representativeness model are shown for (b) Experiments 1b and 2b. Error bars indicate 95% bootstrapped confidence intervals.

**Table 1.** Estimates of a Bayesian mixed-effects regression fit for the following model:  $P(\text{repeat}) \sim 1 + \text{streak length}^2 + \text{streak length} + (1 + \text{streak length}^2 + \text{streak length} \mid \text{participant})$

	Estimate	Estimated error	95% CrI
Experiment 1b			
Intercept	63.82	3.86	[56.34, 71.47]
Streak length <sup>2</sup>	1.58	0.38	[0.82, 2.32]
Streak length	-15.85	2.88	[-21.40, 10.24]
Experiment 3			
Intercept	47.81	1.76	[44.28, 51.25]
Streak length <sup>2</sup>	0.12	0.13	[-0.12, 0.37]
Streak length	-0.79	1.16	[-3.05, 1.47]

Note: Estimates and estimated errors are means and standard deviations of the posterior distributions, respectively. CrI = credible interval.

with terminal streak color  $M$  and solved for the required number of balls with terminal streak color  $X$ :

$$P(\text{ground truth}) = (M + X) / (N + 1), \quad (1)$$

where  $N = 8$  is the number of balls in the sequence. Intuitively, this equation attempts to bring the proportion of the terminal streak color to the ground truth probability by adding  $X$  (which can be positive or negative). We then passed  $X$  through a sigmoid to generate a probability judgment:

$$P(\text{repeat}) = 1 / (1 + \exp(-X)). \quad (2)$$

For point predictions, we computed the proportion of balls with terminal streak color ( $M/8$ ) and compared that to the ground truth. If the proportion is smaller than the ground truth, the model predicts repetition. If the proportion is larger than the ground truth, the model predicts reversal. If the proportion equals the ground truth, the model randomly predicts the next ball color on the basis of the ground truth. As shown in Figures 5a and 5b, the representativeness model was unable to capture the data.

Last, we tested for the nonlinearity we observed in Experiment 1b (the U-shaped data in the first row of Fig. 5b) as an exploratory analysis. We fit a Bayesian mixed-effects model predicting  $P(\text{repeat})$  with terminal streak length, quadratic terminal streak length, and intercept, with random intercept and random slopes for terminal streak length and quadratic terminal streak length grouped by participants (see Table 1). We found a positive effect for the quadratic term ( $b = 1.58$ ), and

the 95% CrI—95% CrI = [0.82, 2.32]—excluded zero. This finding was consistent with the U-shape in Figure 5b.

## Discussion

We failed to observe the gambler's fallacy in participants' probability judgments (Experiment 1a). In contrast, we observed the gambler's fallacy in point predictions (Experiment 1b). However, they were not well explained by a transformation of the probability judgments, suggesting that qualitatively different mechanisms underlie these two types of responses, and the locus of the effect is probably in the decision process rather than the perception of probability.

Our findings contrast with those of Rao and Hastie (2023), who found a small gambler's fallacy in probability judgments. Considering that the only difference between Experiment 1a and their experiment was the sequences—we used IID sequences whereas Rao and Hastie (2023) used non-IID sequences—there is reason to suspect that the effect they found originated from the dependencies in their sequences. We explored this hypothesis further in Experiment 3, in which we replicated Rao and Hastie's (2023) study.

## Experiments 2a and 2b

Experiments 2a and 2b aimed to test an alternative explanation for the discrepancy in probability judgments and point predictions. Because there was no right or wrong prediction to make when the two outcomes were equally likely, it is possible that participants resorted to a nonprobabilistic heuristic to make point predictions. To test whether this was the case, we conducted a new set of experiments with the same setup except that we gave participants IID sequences from a process 60% likely to produce one ball color and 40% likely to produce the other ball color. If, with non-50% IID sequences, we can recover point predictions by sampling from probability judgments, that would be evidence suggesting that participants used the same underlying mechanism for probability judgments and point predictions and that the discrepancy we observed previously was merely an artifact of 50% IID sequences. If instead we still see a deviation between probability judgments and point predictions, that would suggest a more robust difference between the two types of judgments.

## Method

**Participants.** We recruited 150 participants for each experiment via Amazon's MTurk platform. As in Experiments 1a and 1b, participants completed a comprehension check after reading the instructions. Participants



were compensated \$2.50 to complete 18 trials. The experiments were approved by the Harvard Institutional Review Board and preregistered at <https://aspredicted.org/prws-hqh3.pdf>.

**Stimuli.** The stimuli were 18 IID sequences of eight random ball colors generated on the fly on each trial for each participant. Each ball color was independently sampled from a Bernoulli distribution with  $p = .6$  for one color and  $p = .4$  for the other color (randomized for each participant).

**Procedure.** The procedure was the same as Experiments 1a and 1b except that the covered cage contained either 60 blue balls and 40 red balls or 40 blue balls and 60 red balls, in accordance with the ground truth probabilities.

## Results

Because the two ball colors had different ground truth probabilities (60% and 40%), we grouped the trials by the ground truth probability of the terminal ball color and analyzed them separately. In Experiment 2a, for a ground truth probability of 60%, a Lilliefors (Kolmogorov-Smirnov) test for normality indicated that the data were not normally distributed ( $D = 0.12$ ,  $p < .001$ ). A one-sample Wilcoxon signed-rank test indicated that the median was significantly different from 60% ( $Z = -3.65$ ,  $p < .001$ ,  $r = .30$ ). A Bayesian  $t$  test yielded a  $BF_{10}$  of 421.921, meaning that the data were more than 400 times more likely under the alternative hypothesis that the mean would be different from 60%. The median resulting posterior distribution for the effect size ( $\delta$ ) was  $-0.34$ —95% CrI =  $[-0.50, -0.18]$ . For a ground truth probability of 40%, a Lilliefors (Kolmogorov-Smirnov) test for normality indicated that the data were not normally distributed ( $D = 0.10$ ,  $p < .001$ ). A one-sample Wilcoxon signed-rank test indicated that the median was not significantly different from 40% ( $Z = -1.77$ ,  $p = .076$ ,  $r = .14$ ). A Bayesian  $t$  test yielded a  $BF_{10}$  of 0.148, meaning that the data were more than six times more likely under the null hypothesis that the mean would not be different from 40%. The median resulting posterior distribution for the effect size ( $\delta$ ) was  $-0.08$ —95% CrI =  $[-0.24, 0.08]$ . See the second and third rows of Figure 4 (left) for the distribution of  $P(\text{repeat})$ . As before, these effects seemed to be driven by a small percentage of outliers. For trials ending in the color with 60% probability, the mean  $P(\text{repeat})$  was 55.47%, and the median  $P(\text{repeat})$  was 58.25%. For trials ending in the color with 40% probability, the mean  $P(\text{repeat})$  was 39.04%, and the median  $P(\text{repeat})$  was 38.31%. The second and third rows of Figure 5a show the mean and median predictions by terminal streak length.

In Experiment 2b, for a ground truth probability of 60%, a Lilliefors (Kolmogorov-Smirnov) test for normality indicated that the data were not normally distributed ( $D = 0.14$ ,  $p < .001$ ). A one-sample Wilcoxon signed-rank test indicated that the median was not significantly different from 60% ( $Z = -0.95$ ,  $p = .341$ ,  $r = .08$ ). A Bayesian  $t$  test yielded a  $BF_{10}$  of 0.149, meaning that the data were six times more likely under the null hypothesis that the mean would not be different from 60%. The median resulting posterior distribution for the effect size ( $\delta$ ) was  $-0.08$ —95% CrI =  $[-0.24, 0.08]$ . For a ground truth probability of 40%, a Lilliefors (Kolmogorov-Smirnov) test for normality indicated that the data were not normally distributed ( $D = 0.21$ ,  $p < .001$ ). A one-sample Wilcoxon signed-rank test indicated that the median was significantly different from 40% ( $Z = -7.57$ ,  $p < .001$ ,  $r = .62$ ). A Bayesian  $t$  test yielded a  $BF_{10} > 1,000$ , meaning that the data were 1,000 times more likely under the alternative hypothesis that the mean would be different from 40%. The median resulting posterior distribution for the effect size ( $\delta$ ) was  $-0.76$ —95% CrI =  $[-0.95, -0.58]$ . See the second and third rows of Figure 4 (right) for the distribution of  $P(\text{repeat})$ . This effect was primarily driven by very long streak lengths (when all eight balls are in the color with 40% probability; see Fig. 5b). The gambler's fallacy did not show up robustly in the 50% or 60% conditions, presumably because seeing such long streaks of a low-probability event makes people suspicious that the process is not following the described structure.

We again compared the subjective and objective probability models to the point prediction patterns. As shown in the second and third rows of Figure 5b, the point predictions in Experiment 2b were not well explained by a transformation of the probability judgments in Experiment 2a or the ground truth probabilities. In particular, the subjective model was in the opposite direction of the data in the 40% condition as well as the objective model in the 60% condition. The representativeness model also could not explain the data in either experiment. Finally, the thresholding model seemed closer to the data in the 60% condition visually (see Fig. S1) but not in the 40% condition. However, Bayesian  $t$  tests comparing participants'  $P(\text{repeat})$  in Experiment 2b and the thresholding model's point predictions based on participants' probability judgments in Experiment 2a showed that the thresholding model could not recover the point prediction data in either condition. The 60% condition yielded a  $BF_{10}$  of 2.097, meaning that the data were two times more likely under the alternative hypothesis that the mean of participants'  $P(\text{repeat})$  would be different from the model's predicted  $P(\text{repeat})$ . The median resulting posterior distribution for the effect size ( $\delta$ ) was 0.27—95%

CrI = [0.05, 0.50]. The 40% condition yielded a  $BF_{10} > 1,000$ , meaning that the data were more than 1,000 times more likely under the alternative hypothesis that the mean of participants'  $P(\text{repeat})$  would be different from the model's predicted  $P(\text{repeat})$ . The median resulting posterior distribution for the effect size ( $\delta$ ) was  $-0.65$ —95% CrI =  $[-0.88, -0.42]$ .

## Discussion

Experiments 2a and 2b demonstrated the robustness of our findings in Experiments 1a and 1b by introducing asymmetric generators. As in the previous experiments, we did not find the gambler's fallacy consistently, and the effects we did find were still numerically small. When we did see strong effects for probability judgments in the 40% condition, they appeared only at very long streaks, and they also did not show up consistently in the 50% or 60% conditions. Last, we still were not able to reconstruct point predictions from sampling from probability judgments. Thus, our conclusions from Experiments 1a and 1b appear to hold rather broadly.

## Experiment 3

Experiment 3 replicated an experiment reported by Rao and Hastie (2023). Our goal was to examine the robustness of their findings and further investigate why they found the gambler's fallacy in probabilistic predictions, whereas we did not. We hypothesized that the sequences used in Rao and Hastie (2023), which we refer to as "RH sequences," had statistics different from IID sequences (because they were filtered by Rao and Hastie) and that participants might have been sensitive to that difference. Critically, all of our experiments used truly IID sequences. Here we explored the effects of using RH sequences.

## Method

**Participants.** We recruited 150 participants via Amazon's MTurk platform. As in the experiments described above, participants completed a comprehension check after reading the instructions. Participants were compensated \$2.50 to complete 18 trials. The experiment was approved by the Harvard Institutional Review Board and preregistered at <https://aspredicted.org/khp6-hkz8.pdf>.

**Stimuli.** The stimuli were identical to Rao and Hastie (2023). The stimuli contained 18 sequences of eight ball colors. Twelve of the sequences ended with streak length of 1. The remaining six sequences ended with a streak length of 2, 3, 4, 5, 6, and 7. All of the sequences were drawn from a pool of designed sequences for each

participant. The order was shuffled for each participant, but the first sequence always ended with streak length of 1.

**Procedure.** The procedure was identical to Experiments 1a and 1b.

## Results

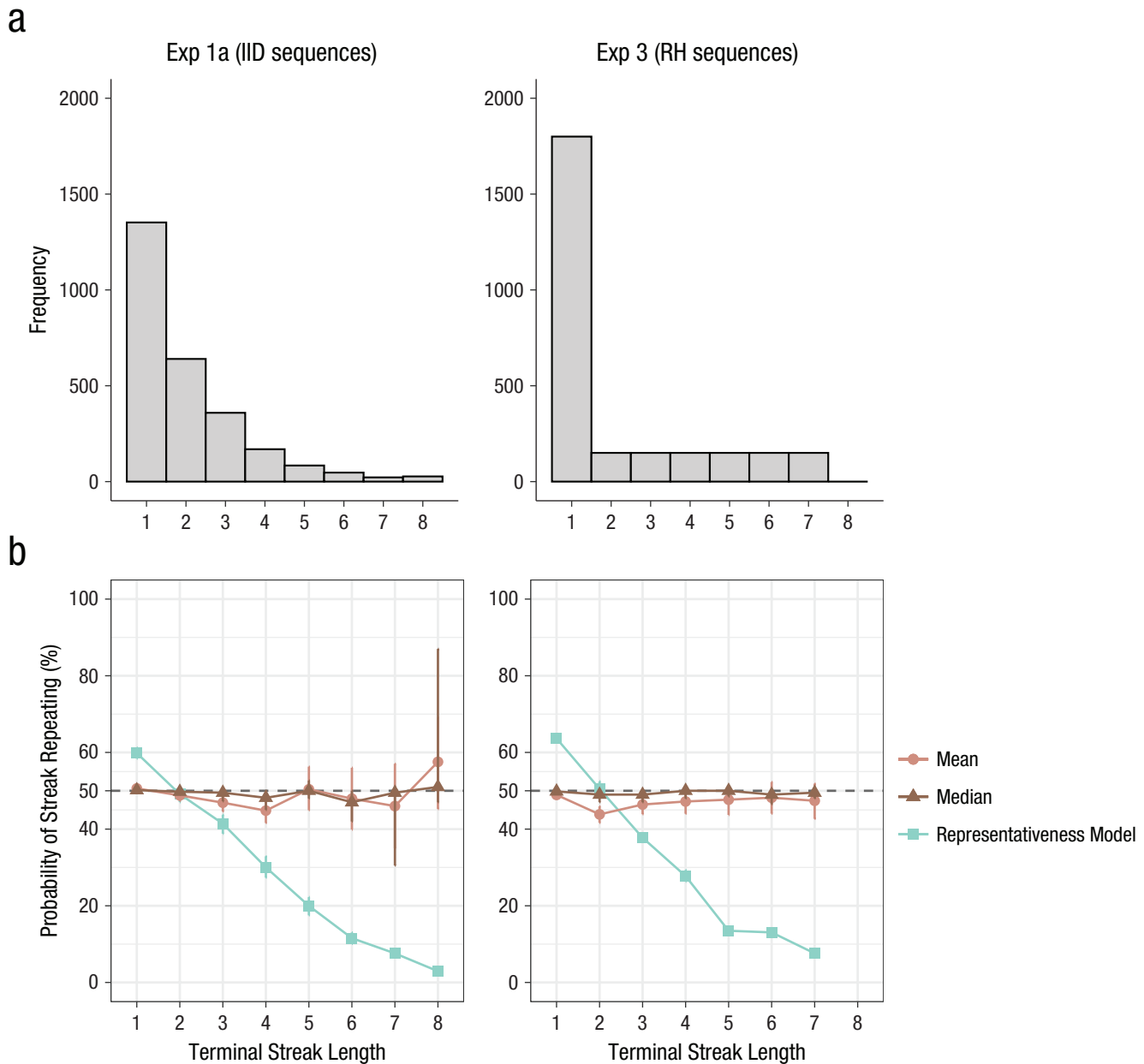
Experiment 3 was a direct comparison with Experiment 1a because they differed only in the sequence-generation process. As shown in Figure 6a, there were meaningful differences in the distribution of the terminal streak length between these two experiments: The IID sequences used in Experiment 1a tended to show an exponential distribution with a more gradual decrease in frequency as the streak lengths increased, whereas the RH sequences used in Experiment 3 consisted of a uniform distribution of streak lengths 2 to 7, with streak length of 1 appearing much more frequently than truly IID sequences.

Figure 6b juxtaposes the mean and median predictions by terminal streak length in Experiments 1a and 3. A two-sample Wilcoxon test showed that the median in Experiment 3 was significantly different from the median in Experiment 1a ( $Z = -2.25$ ,  $p = .025$ ,  $r = .13$ ).

A Lilliefors (Kolmogorov-Smirnov) test for normality indicated that the data in Experiment 3 were not normally distributed ( $D = 0.12$ ,  $p < .001$ ). A one-sample Wilcoxon signed-rank test indicated that the median was significantly different from 50% ( $Z = -4.02$ ,  $p < .001$ ,  $r = .33$ ). A Bayesian  $t$  test yielded a  $BF_{10}$  of 273.609, meaning that the data were more than 200 times more likely under the alternative hypothesis that the mean would be different from 50%. The median resulting posterior distribution for the effect size ( $\delta$ ) was  $-0.33$ —95% CrI =  $[-0.50, -0.17]$ . However, this effect seemed to be driven by a small percentage of participants. Only 7.33% of participants'  $P(\text{repeat})$  was less than 40%. As shown by Figure 6b (right), it also seemed to be driven by a few streak lengths—primarily streak lengths 2 and 3. To characterize this, we fit a Bayesian mixed-effects model predicting  $P(\text{repeat})$  with terminal streak length, quadratic terminal streak length, and intercept, with random intercept and random slopes for terminal streak length and quadratic terminal streak length grouped by participants (see Table 1). The coefficient of the quadratic term was positive ( $b = 0.12$ ), but the 95% CrI—95% CrI =  $[-0.12, 0.37]$ , did not exclude zero.

## Discussion

In Experiment 3, we replicated the small gambler's fallacy for the probabilistic predictions reported in Rao



**Fig. 6.** Comparison between Experiments 1a and 3. The (a) distribution of terminal streak lengths in the stimuli in Experiments 1a and 3 show an exponential distribution with a more gradual decrease in frequency as the streak lengths increase for the IID sequences and a uniform distribution of streak lengths 2 to 7 for the RH sequences, with a streak length of 1 appearing much more frequently than truly IID sequences. The (b) probability that a streak will repeat for different terminal streak lengths is also shown. Error bars indicate bootstrapped 95% confidence intervals. IID = independent and identically distributed; RH = Rao and Hastie (2023).

and Hastie (2023). A comparison between Experiments 1a and 3 revealed that Rao and Hastie's (2023) sequences did not have the same statistics as IID sequences; participants were clearly sensitive to this difference even though they were told in both experiments that the sequences were truly random. Still, the effects we observed were small, holding only for a small subset of participants and streak lengths. The effects likely

resulted from nonrandomness in the sequences because we did not find them in Experiment 1a.

## General Discussion

In the current work, we tested the gambler's fallacy through a series of experiments using IID sequences. Overall, we failed to observe a robust gambler's fallacy

in participants' probabilistic predictions (Experiments 1a and 2a). By contrast, we observed the gambler's fallacy when participants made point predictions (Experiments 1b and 2b) or when they were given correlated processes (Experiment 3), in which case the effect might be better characterized as a "pseudo-gambler's fallacy."

Our findings challenge the idea, found throughout the modeling literature (and reviewed earlier on in this article), that the gambler's fallacy arises from probabilistic reasoning. We were not able to reconstruct the fallacy apparent in point predictions from the probabilistic predictions using a simple sampling model. According to this model, participants draw samples sequentially and use this "Monte Carlo" estimate to generate a point prediction by an optional stopping rule. Although there is much to recommend this kind of model from other quarters of cognitive science (Griffiths et al., 2012; Sanborn, 2017), there clearly must be some other process giving rise to the fallacy in point predictions.

The disconnection between probabilistic predictions and point predictions also rules out a few alternative explanations. We explored a representativeness model based on the representativeness heuristic (Kahneman & Tversky, 1972; Tversky & Kahneman, 1974); it made similar predictions for probability judgments and point predictions (see Fig. 5), suggesting that the representativeness heuristic alone cannot explain the two judgments. Likewise, a conservatism bias (the tendency toward 50%; Edwards, 1968) or noisy cognition models (Costello & Watts, 2014, 2016; Enke & Graeber, 2023; Xiang et al., 2021) would not be able to explain the discrepancies between the probabilistic and point predictions we observed. Critically, if point predictions are functions of probabilistic predictions, then they should reflect the same biases measured in probabilistic predictions. Our results show that point predictions cannot be easily reconstructed from probabilistic predictions regardless of what mechanism underlies the probabilistic predictions. There is conceivably some other mapping from probabilistic to point predictions, but existing models do not specify what this might look like. Our results do not rule out the possibility that the probability judgments participants reported do not reflect their internal model of probabilistic reasoning because people have difficulty reporting precise probability judgments, especially with an unfamiliar task and an unfamiliar elicitation mechanism (Gigerenzer, 1991, 1993). However, the same task and elicitation mechanism were used in Rao and Hastie (2023), who found that when participants were uncertain whether the ground truth was 25%, 50%, or 75%, they reported probability judgments that significantly deviated from 50% as the streak length increased, in line with Bayesian posterior probabilities. It is therefore reasonable to infer

that participants in this task are capable of making probability judgments that reflect their internal model.

The most parsimonious explanation, then, is that point prediction does not rely on probabilistic reasoning. What might a viable alternative model look like? Generally speaking, what is needed is a model that generates point predictions from a function of outcome history. Critically, this function does not require computational access to probabilistic predictions, which might be generated from a completely separate process. An intriguing possibility is that cognitive heuristics such as representativeness or the law of small numbers arise from naturalistic sequence statistics combined with an information or computational bottleneck. Some suggestive hints in this direction have been glimpsed in large language models (Castello et al., 2024; Suri et al., 2024). Other heuristics might also be involved because we showed that the representativeness heuristic itself could not capture the point prediction data quantitatively, and in some cases, not even qualitatively (the U-shape in Experiment 1b).

Our findings also have implications for experiments that test the gambler's fallacy and judgment and decision-making in general. The comparison between Experiment 1a (which used IID sequences) and Experiment 3 (which used non-IID sequences) revealed that non-IID sequences did not have the same statistics as IID sequences, and participants responded differently to non-IID sequences despite being told that they were IID. This suggests that participants were sensitive to the statistical patterns of the stimuli they were presented with. Experiments that use deception should be mindful of its effects on participants' responses because they may infer that the stimuli are not generated by the described process. The current research is potentially limited to MTurk workers; therefore, future work is needed to comprehensively test this idea.

In sum, we have provided evidence that the gambler's fallacy likely originates at the decision stage rather than in probabilistic reasoning. Our findings challenge the existing theories of the gambler's fallacy, which typically (if not exclusively) explain the fallacy as arising from probabilistic reasoning, whether from irrational errors or from rational probabilistic inference. Emerging theories of the fallacy may need to explain it without relying on probabilistic predictions.

## Transparency

*Action Editor:* Gabriele Paolacci

*Editor:* Simine Vazire

*Author Contributions*

**Yang Xiang:** Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Software; Validation; Visualization; Writing – original draft; Writing – review & editing.



**Kevin Dorst:** Conceptualization; Methodology; Supervision; Writing – review & editing.

**Samuel J. Gershman:** Conceptualization; Funding acquisition; Methodology; Supervision; Writing – review & editing.

#### Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

#### Funding

This work was supported National Science Foundation Grant DRL-2024462 (to S. J. Gershman).

#### Artificial Intelligence

No AI-assisted technologies were used in this research or the creation of this article.

#### Ethics


This research received approval from the Harvard Institutional Review Board (Protocol No. IRB15-2048).


#### Open Practices

Experiments 1a and 1b disclosures. Preregistration: The research questions, method, and analyses were preregistered (<https://aspredicted.org/2frb-zshx.pdf>) on June 29, 2024, prior to data collection, which began on July 1, 2024. There were no deviations from the preregistration. Materials: All study materials are publicly available (<https://osf.io/67nsd>). Data: All primary data are publicly available (<https://osf.io/67nsd>). Analysis scripts: All analysis scripts are publicly available (<https://osf.io/67nsd>). Experiments 2a and 2b disclosures. Preregistration: The research questions, method, and analyses were preregistered (<https://aspredicted.org/prws-hqh3.pdf>) on July 17, 2024, prior to data collection, which began on July 23, 2024. There were no deviations from the preregistration. Materials: All study materials are publicly available (<https://osf.io/67nsd>). Data: All primary data are publicly available (<https://osf.io/67nsd>). Analysis scripts: All analysis scripts are publicly available (<https://osf.io/67nsd.io/67nsd>). Experiment 3 disclosures. Preregistration: The research questions, method, and analyses were preregistered (<https://aspredicted.org/khp6-hkz8.pdf>) on July 9, 2024, prior to data collection, which began later that same day. There were no deviations from the preregistration. Materials: All study materials are publicly available (<https://osf.io/67nsd>). Data: All primary data are publicly available (<https://osf.io/67nsd>). Analysis scripts: All analysis scripts are publicly available (<https://osf.io/67nsd>).

#### ORCID iDs

Yang Xiang  <https://orcid.org/0000-0001-9083-3265>

Kevin Dorst  <https://orcid.org/0000-0003-3982-3242>

Samuel J. Gershman  <https://orcid.org/0000-0002-6546-3298>

#### Acknowledgments

We are grateful to Dean Eckles, Arthur Prat-Carrabin, and Eric Bigelow for helpful discussions.

#### Supplemental Material

Additional supporting information can be found at <http://journals.sagepub.com/doi/suppl/10.1177/09567976251344570>

#### Note

1. In the early literature on probability learning, starting with Jarvik (1951), the gambler's fallacy is known as the "negative recency effect" (for a review, see Jones, 1971).

#### References

- Agliari, A., Hommes, C. H., & Pecora, N. (2016). Path dependent coordination of expectations in asset pricing experiments: A behavioral explanation. *Journal of Economic Behavior & Organization*, 121, 15–28.
- Asparouhova, E., Hertzel, M., & Lemmon, M. (2009). Inference from streaks in random outcomes: Experimental evidence on beliefs in regime shifting and the law of small numbers. *Management Science*, 55(11), 1766–1782.
- Ayton, P., & Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Memory & Cognition*, 32, 1369–1378.
- Barberis, N., Shleifer, A., & Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics*, 49(3), 307–343.
- Barron, G., & Leider, S. (2010). The role of experience in the gambler's fallacy. *Journal of Behavioral Decision Making*, 23(1), 117–129.
- Bloomfield, R., & Hales, J. (2002). Predicting the next step of a random walk: Experimental evidence of regime-shifting beliefs. *Journal of Financial Economics*, 65(3), 397–414.
- Castello, M., Pantana, G., & Torre, I. (2024). Examining cognitive biases in ChatGPT 3.5 and ChatGPT 4 through human evaluation and linguistic comparison. In R. Knowles, A. Eriguchi, & S. Goel (Eds.), *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)* (pp. 250–260). Association for Machine Translation in the Americas.
- Clotfelter, C. T., & Cook, P. J. (1993). The "gambler's fallacy" in lottery play. *Management Science*, 39(12), 1521–1525.
- Costello, F., & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review*, 121(3), 463–480.
- Costello, F., & Watts, P. (2016). People's conditional probability judgments follow probability theory (plus noise). *Cognitive Psychology*, 89, 106–133.
- Dorst, K. (2024). *Bayesians commit the gambler's fallacy*. SSRN. <https://doi.org/10.2139/ssrn.4683064>
- Edwards, W. (1968). *Conservatism in human information processing*. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp. 17–52). John Wiley & Sons.
- Enke, B., & Graeber, T. (2023). Cognitive uncertainty. *The Quarterly Journal of Economics*, 138(4), 2021–2067.
- Farmer, G., Warren, P., & Hahn, U. (2017). Who "believes" in the gambler's fallacy and why? *Journal of Experimental Psychology: General*, 146(1), 63–76.

- Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases." *European Review of Social Psychology*, 2(1), 83–115.
- Gigerenzer, G. (1993). The bounded rationality of probabilistic mental models. In K. I. Manktelow & D. E. Over (Eds.). *Rationality: Psychological and philosophical perspectives* (pp. 284–313). Routledge.
- Glazer, J., & Rubinstein, A. (2024). Making predictions based on data: Holistic and atomistic procedures. *Journal of Economic Theory*, 216, Article 105791. <https://doi.org/10.1016/j.jet.2023.105791>
- Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, 21(4), 263–268.
- Hahn, U., & Warren, P. A. (2009). Perceptions of randomness: Why three heads are better than four. *Psychological Review*, 116(2), 454.
- Icard, T. (2016). Subjective probability as sampling propensity. *Review of Philosophy and Psychology*, 7, 863–903.
- Jarvik, M. (1951). Probability learning and a negative recency effect in the serial anticipation of alternative symbols. *Journal of Experimental Psychology*, 41(4), 291–297.
- Jones, M. R. (1971). From probability learning to sequential processing: A critical review. *Psychological Bulletin*, 76(3), 153–185.
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3(3), 430–454.
- Kong, Q., Granic, G. D., Lambert, N. S., & Teo, C. P. (2020). Judgment error in lottery play: When the hot hand meets the gambler's fallacy. *Management Science*, 66(2), 844–862.
- Marquis de Laplace, P. S. (1902). *A philosophical essay on probabilities*. John Wiley & Sons.
- McClelland, G. H., & Hackenberg, B. H. (1978). Subjective probabilities for sex of next child: Us college students and Philippine villagers. *Journal of Population*, 1(2), 132–147.
- Miller, J. B., & Sanjurjo, A. (2018). *How experience confirms the gambler's fallacy when sample size is neglected*. OSF. <https://doi.org/10.31219/osf.io/m5xsk>
- Misirlişoy, E., & Haggard, P. (2014). Asymmetric predictability and cognitive competition in football penalty shootouts. *Current Biology*, 24(16), 1918–1922.
- Rabin, M. (2002). Inference by believers in the law of small numbers. *The Quarterly Journal of Economics*, 117(3), 775–816.
- Rabin, M., & Vayanos, D. (2010). The gambler's and hot-hand fallacies: Theory and applications. *The Review of Economic Studies*, 77(2), 730–778.
- Rao, K., & Hastie, R. (2023). Predicting outcomes in a sequence of binary events: Belief updating and gambler's fallacy reasoning. *Cognitive Science*, 47(1), Article e13211. <https://doi.org/10.1111/cogs.13211>
- Roney, C. J., & Sansone, N. (2015). Explaining the gambler's fallacy: Testing a gestalt explanation versus the "law of small numbers." *Thinking & Reasoning*, 21(2), 193–205.
- Roney, C. J., & Trick, L. M. (2003). Grouping and gambling: A gestalt approach to understanding the gambler's fallacy. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 57(2), 69–75.
- Sanborn, A. N. (2017). Types of approximation for probabilistic cognition: Sampling and variational. *Brain and Cognition*, 112, 98–101.
- Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, 20(12), 883–893.
- Suetens, S., & Tyran, J.-R. (2012). The gambler's fallacy and gender. *Journal of Economic Behavior & Organization*, 83(1), 118–124.
- Sundali, J., & Croson, R. (2006). Biases in casino betting: The hot hand and the gambler's fallacy. *Judgment and Decision Making*, 1(1), 1–12.
- Suri, G., Slater, L. R., Ziaee, A., & Nguyen, M. (2024). Do large language models show decision heuristics similar to humans? A case study using GPT-3.5. *Journal of Experimental Psychology: General*, 153(4), 1066–1075.
- Tergiman, C. (2015). Institution design and public good provision: An experimental study of the vote of confidence procedure. *Experimental Economics*, 18, 697–717.
- Terrell, D. (1994). A test of the gambler's fallacy: Evidence from pari-mutuel games. *Journal of Risk and Uncertainty*, 8, 309–317.
- Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 76(2), 105–110.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, 185(4157), 1124–1131.
- Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, 38(4), 599–637.
- Weiss, D. J., & Shanteau, J. (2004). The vice of consensus and the virtue of consistency. In K. Smith, J. Shanteau, & P. Johnson (Eds.), *Psychological investigations of competence in decision making* (pp. 226–240). Cambridge University Press.
- Weitzel, U., & Rosenkranz, S. (2016). *Randomness and the madness of crowds*. Springer.
- Xiang, Y., Graeber, T., Enke, B., & Gershman, S. J. (2021). Confidence and central tendency in perceptual judgment. *Attention, Perception, & Psychophysics*, 83, 3024–3034.
- Zhu, J.-Q., Sundh, J., Spicer, J., Chater, N., & Sanborn, A. N. (2024). The autocorrelated Bayesian sampler: A rational process for probability judgments, estimates, confidence intervals, choices, confidence judgments, and response times. *Psychological Review*, 131(2), 456–493.