First draft: November 19th, 2025
This draft: December 1st , 2025

Commentary on Xiang et al. (2025): The Gambler's Fallacy Fallacy Fallacy

Daniel Banki
Esade Business School
bankidaniel@gmail.com

Uri Simonsohn
Esade Business School
urisohn@gmail.com

Abstract.
In contrast to the results and conclusions by Xiang et al. (2025), reanalyzing their data we find significant and substantive evidence of gambler's fallacy beliefs. Our results differ from theirs because our analytical approach differs from theirs. While they average probability estimates across judgments, we count probability estimates exhibiting the gambler's fallacy. For example, after a streak of length four, 57% of responses deem a streak as overly likely to end, compared to 31% that do the opposite. Moreover, Xiang et al. report that the *median* participant doesn't exhibit the gambler's fallacy, while we find that a substantial *minority* of participants do exhibit it.

Reproduction Package: https://ResearchBox.org/5013 (Use Code: TZZRVQ)

Xiang et al. (2025) write: "We failed to observe the gambler's fallacy in participants' probability judgments" (p. 8). We reanalyze their posted data and, in contrast, do observe the gambler's fallacy in those same probability judgments. Our results differ from theirs because our analytical approach differs from theirs.

In their studies, participants saw 8 randomly drawn blue and red balls and gave a probability for the 9th ball being red. Each participant did this 18 times. Xiang et al.'s main analysis, which we will refer to as the "aggregating method", computed the average probability each participant gave across those 18 judgments, and compared the median participant's average to the true probability (e.g., to 50% in Study 1).

In contrast, we propose the "counting method", where we count how many probability estimates are greater than vs. lower than 50%, or how many participants behave in line with the gambler's fallacy.[1]

To see why results may change when counting instead of aggregating, imagine a participant estimating the probability of a red ball following streaks of 1, 2, 3, 4, and 5 blue balls in a row respectively, answering 24%, 55%, 56%, 57% and 58%, respectively. The *average* probability is 50%, no gambler's fallacy. But 4 out of 5 estimates are in line with the gambler's fallacy. In terms of counting participants, imagine that 40% of them show the gambler's fallacy, 55% behave rationally (predict 50:50), and 5% exhibit the hot hand. The *median* participant is rational, but there is a sizeable minority of participants who show the gambler's fallacy.

[1] In an experimental design where every participant gives a single response, the two implementations of the counting method are equivalent: computing the share of responses/participants above, at, and below 50%. But when individual participants give multiple responses, as they do here, it is ambiguous how to code a participant who gives some but not all responses above 50%. For robustness, we carry out both prediction-level and participant-level analyses.

The counting method more directly answers what we see as the research question of interest: how common is the gambler's fallacy? The aggregating method, in contrast, answers a proxy and less interpretable question: does the average (or median) response show the gambler's fallacy? The average response need not be representative of the majority of responses (Bolger et al., 2019; Chen et al., 2021; Hershey & Schoemaker, 1980; Hutchinson et al., 2000); indeed, it need not be representative of even a minority of responses. In some studies, nobody is average. Counting has another advantage: greater statistical power to detect the gambler's fallacy. In scenarios we consider plausible, the power advantage surpasses 40 percentage points (see Supplement 1).

Despite its conceptual and statistical superiority, the counting method is not something researchers commonly do, neither when studying the gambler's fallacy nor when studying other behavioral phenomena. One reason for that is that it is not commonly applicable. In most studies, a single response cannot be categorized as supporting or contradicting a hypothesis. For example, if a participant predicts "red" after seeing four blue balls, do they exhibit the gambler's fallacy? We don't know. They may correctly think the probability of "red" is 50% and *randomly* choose "red". If we cannot categorize a single response as supporting/contradicting a hypothesis, we cannot count how many of them do. We similarly cannot judge whether a single estimate of the length of the river Nile is influenced by an anchor, nor whether a single valuation of a Cornell mug is influenced by being a seller vs. buyer. But in this gambler's fallacy experiment we can evaluate individual responses because participants provide a probability estimate for an outcome whose true probability is known.

Methods

The gambler's fallacy is the belief that streaks produced by random sequences are overly likely to come an end (Marquis de Laplace, 1902, pp. 161-162). [2] It is a phenomenon generally attributed to the "representativeness heuristic", where people's *beliefs about probabilities* are impacted by how well a sample represents prototypical characteristics of the population (Kahneman & Tversky, 1972). Xiang et al. (2025) invite us to reject this explanation of the gambler's fallacy, proposing beliefs about probabilities "do not play a role [in it]" (p. 12).

Gambler's fallacy studies typically involve people predicting the next outcome after observing a set of random outcomes (Clotfelter & Cook, 1993; Croson & Sundali, 2005; Terrell, 1994) or people attempting to generate random sequences (Rabin, 2002; Rapoport & Budescu, 1992). In contrast, Xiang et al. (2025) key results are obtained from studies in which they ask participants about the probability of a given outcome (this design is based on earlier work by Rao and Hastie (2023)).

Original design and analysis

As mentioned above, in the studies by Xiang et al. (2025), participants were presented with 18 sequences of 8 randomly drawn blue or red balls and asked about the expected color of the 9th ball. The gambler's fallacy is present if participants believe the next ball is more likely to end rather than to continue a streak (e.g., after three blue balls, a red ball is perceived as more likely than a blue one). Participants were assigned to either (i) predicting the color of the next ball (answering "blue" or "red"), or (ii) indicating the probability that the next ball will be red (on a

[2] Marquis de Laplace (1902) is often credited as the first to discuss the gambler's fallacy. He writes, "It is, for example, very improbable that at the play of heads and tails one will throw heads ten times in succession ... when it has happened nine times, [it] leads us to **believe** that at the tenth throw tails will be thrown" (**bold** added; p. 162)

slider where (only) the extremes, 0% and 100%, were labeled).[3] The studies differ in the baseline proportion of red to blue balls. In Studies 1 and 3, the proportions were 50:50; in Study 2, the proportions were either 60:40 or 40:60.

In each study, the dependent variable is categorical ("red" or "blue") in one condition and numerical (probability of red) in the other. In order to analyze both conditions with the same analytical approach, the authors converted the categorical predictions to imputed probabilities. Specifically, answers of "red" were converted to "100% chance of red" and answers of "blue" to "0% chance of red". In both conditions, they averaged each participant’s 18 probabilities, computed the median across participants, and compared this median to the true probability (e.g., 50% in Study 1), with a one-sample Wilcoxon test.

They report other analyses that focus on the difference between conditions rather than on whether there is evidence of the gambler's fallacy in the probability condition. The between-condition analyses are secondary for our purposes and thus not discussed here (the General Discussion focuses on the interpretability of differences of results across the two elicitation methods).

Original Results

In Study 1, participants in the prediction condition, where they indicated whether the next ball would be "red" or "blue", had a median imputed 'probability' of the streak ending of

[3] Xiang et al. run these two arms as separate studies (e.g., "Study 1a" and "Study 1b"), without random assignment. But, for ease of exposition, we refer to them as conditions of a single study (e.g., "Study 1").

Mdn = 61%; significantly different from 50%, Z = 5.56, $p < .001$.[4,5] In contrast, in the condition where participants directly provided a probability of red, the median was not different from 50%, *Mdn* = 50%, Z = 1.03, $p = .304$.

In a follow-up analysis (see their Figure 5), Xiang et al. (2025) analyze results separately for streaks of each possible length between 1 and 8 (e.g., probability of red after 1 red, after 2 reds, etc.). The gambler's fallacy predicts that as a streak gets longer, the perceived probability that the streak ends increases. They find evidence consistent with this prediction when participants predict "blue" vs "red", but not when they indicate the probability of red.

New analyses: counting instead of averaging

We obtained the data posted by the authors (https://osf.io/hya8z) and with our own code successfully reproduced their key results. We then implemented the counting method by coding each probability assessment as being (i) above, (ii) below, or (iii) equal to the true probability (e.g., 50% in Study 1). We did this for each of the n = 150 participants x 18 responses, N = 2700 observations (per study). Figure 1 plots the proportions of responses above and below the true probability, by streak length, for Studies 1 & 2.[6] We see that for 'streaks' of length 1 (about half of all the data) there is no clear difference but that a gap arises as the streak length increases. For example, for streaks of length 4, a majority of responses (57%) deemed the streak overly likely to end, and a meaningfully smaller proportion of responses (31%) deemed streaks overly likely to continue (the remaining 12% of responses provided the normative 50% answer; we believe this is

[4] The original paper does not report the median values, only their significance level against the null of 50%. The medians reported here were computed by us.

[5] Conducting the Wilcoxon test ourselves, we obtained Z = 6.04 rather than Z = 5.56. We do reproduce the result for the other condition in Study 1: Z = 1.03.

[6] Study 3 involves non-independent draws and shows evidence of the gambler's fallacy even in Xiang et al.'s original analysis. We did not reanalyze it.

an artifactually small proportion caused by participants having to provide responses on an unlabeled slider (see footnote for details).[7]

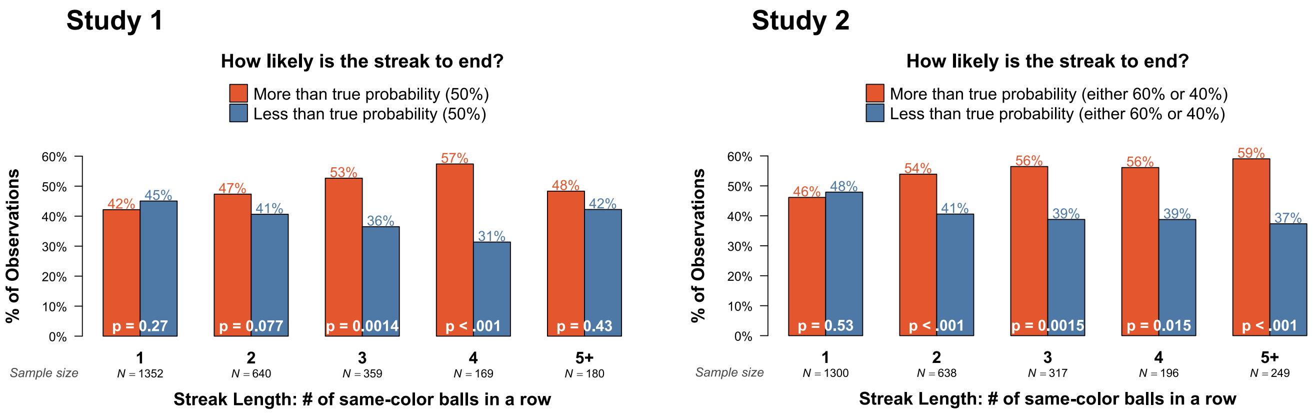

Figure 1. *In line with the gambler's fallacy, overestimates of the probability that the streak ends are more common than underestimates.*

Figure 1 shows results with responses (N = 2700 per study) as the unit of analysis. Computing the proportion of *participants* exhibiting the gambler's fallacy is less straightforward because it requires defining when a *set* of responses by a participant is and is not consistent with the gambler's fallacy. One approach is to compute, for each participant, the average probability that the streak will end when the streak is of length 1 or 2 vs. when it is 3 or 4. A participant is defined to behave in line with the gambler's fallacy when their average probability for lengths of 3 or 4 is larger. In Study 1, we find that this is true of 61.5% of participants, significantly above the 50% expected under the null that streak length does not have an effect ($\chi^2(1) = 7.36$, $p = .007$).

[7] We conjectured, and verified by running their posted experiment software on our own computers, that it is difficult to provide a precise response. Someone wanting to select 50% does not know if the slider is exactly at 50% (getting 60% is even more difficult because one cannot use the button in the middle of the screen as a point of reference for where the desired value may be). Moreover, we found the slider task more tedious than the pick-a-color task, especially in the second half of the 18 rounds. We were puzzled by the high frequency of responses at exactly 60% in Study 2. We then realized that if one right-clicks the slider, the handle appears at exactly that value.

In Study 2, the estimate is 64.4%, also significantly above 50% ($\chi^2(1) = 11.51$, $p < .001$).[8] As a benchmark, in the condition where participants predict "blue" vs "red", rather than provide a probability, these proportions are somewhat higher: 67.8% in Study 1 and 66.2% in Study 2.

General discussion

In contrast to the results and conclusions by Xiang et al. (2025), reanalyzing their data we find significant and substantive evidence of gambler's fallacy beliefs in conditions where participants provide the probability that a streak will end. By some metrics, the gambler's fallacy is smaller when participants give probability estimates rather than predict the next outcome. One interpretation is that in line with Xiang et al.’s general conjecture, (some of) the gambler's fallacy effect that has been documented with predictions does not follow from beliefs. This implies the gambler's fallacy is overestimated in prediction experiments. Another possibility is that the gambler's fallacy is underestimated in probability elicitation experiments. We believe this is likely the case here, due to at least three reasons, each sufficient to challenge the interpretability of differences in results across methods.

First, in Xiang et al.'s experiments, probabilities were elicited with an unlabeled slider scale that makes it difficult to provide precise estimates, introducing more measurement error and thus more attenuation bias in the probability vs. prediction condition (see footnote 7). Second, while the prediction condition involved a forced choice between red and blue, the probability condition allowed answering "50:50", effectively a no-choice option, which may impact decisions on its own

[8] As a robustness check to comparing streaks of 1 or 2, to streaks of 3 or 4, we compared streaks of length 1 to streaks of length 2 or longer; when defined this way, 60.7% & and 67.3% of participants gave a higher probability to longer streaks ending, significantly above the null of 50% of participants doing so (*p*=.011, and p<.001 respectively).

(Dhar, 1997; Dhar & Simonson, 2003; Parker & Schrift, 2011). Third, probabilities are harder to understand than predictions; nobody is confused about what "blue ball" means, but many are confused by what probability means, e.g., people often use 50:50 to mean something other than the probability of an outcome being 50% (Fischhoff & Bruine De Bruin, 1999). Interpreting differences in results obtained with different elicitation mechanisms requires establishing that they have comparable levels of noise and ruling out that other differences in design have unintended effects.

References

Bolger, N., Zee, K. S., Rossignac-Milon, M., & Hassin, R. R. (2019). Causal processes in psychology are heterogeneous. *Journal of Experimental Psychology: General*, *148*(4), 601.

Chen, M., Regenwetter, M., & Davis-Stober, C. P. (2021). Collective choice may tell nothing about anyone's individual preferences. *Decision Analysis*, *18*(1), 1-24.

Clotfelter, C. T., & Cook, P. J. (1993). Notes: The "gambler's fallacy" in lottery play. *Management Science*, *39*(12), 1521-1525.

Croson, R., & Sundali, J. (2005). The gambler's fallacy and the hot hand: Empirical data from casinos. *Journal of risk and uncertainty*, *30*(3), 195-209.

Dhar, R. (1997). Consumer preference for a no-choice option. *Journal of consumer research*, *24*(2), 215-231.

Dhar, R., & Simonson, I. (2003). The effect of forced choice on choice. *Journal of marketing research*, *40*(2), 146-160.

Fischhoff, B., & Bruine De Bruin, W. (1999). Fifty–fifty= 50%? *Journal of Behavioral Decision Making*, *12*(2), 149-163.

Hershey, J. C., & Schoemaker, P. J. H. (1980). Prospect theory's reflection hypothesis: A critical examination. *Organizational Behavior and Human Performance*, *25*(3), 395–418.

Hutchinson, J. W., Kamakura, W. A., & Lynch, J. G. J. (2000). Unobserved heterogeneity as an alternative explanation for "reversal" effects in behavioral research. *Journal of consumer research*, *27*(3), 324-344.

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

Marquis de Laplace, P. S. (1902). *A philosophical essay on probabilities*. Wiley.

Parker, J. R., & Schrift, R. Y. (2011). Rejectable choice sets: How seemingly irrelevant no-choice options affect consumer decision processes. *Journal of marketing research*, *48*(5), 840-854.

Rabin, M. (2002). Inference by believers in the law of small numbers. *The Quarterly Journal of Economics*, *117*(3), 775-816.

Rao, K., & Hastie, R. (2023). Predicting outcomes in a sequence of binary events: belief updating and gambler's fallacy reasoning. *Cognitive Science*, *47*(1), e13211.

Rapoport, A., & Budescu, D. V. (1992). Generation of random series in two-person strictly competitive games. *Journal of Experimental Psychology: General*, *121*(3), 352.

Terrell, D. (1994). A test of the gambler's fallacy: Evidence from pari-mutuel games. *Journal of risk and uncertainty*, *8*(3), 309-317.

Xiang, Y., Dorst, K., & Gershman, S. J. (2025). On the Robustness and Provenance of the Gambler's Fallacy. *Psychological Science*, *36*(6), 451-464.