Judgment and Decision Making, Vol. 4, No. 5, August 2009, pp. 326–334

The retrospective gambler's fallacy: Unlikely events, constructing the past, and multiple universes

Daniel M. Oppenheimer*
Princeton University

Benoît Monin
Stanford University

Abstract

The gambler's fallacy (Tune, 1964) refers to the belief that a streak is more likely to end than chance would dictate. In three studies, participants exhibited a *retrospective gambler's fallacy* (RGF) in which an event that seems rare appears to come from a longer sequence than an event that seems more common. Study 1 demonstrates this bias for streaks, while Study 2 does so with single rare events and shows that the appearance of rarity is more important than actual rarity. Study 3 extends these findings from abstract gambling domains into real world domains to demonstrate the generalizability of the effects. The RGF follows from the law of small numbers (Tversky & Kahneman, 1971) and has many applications, from perceptions of the social world to philosophical debates about the existence of multiple universes.

Keywords: gambler's fallacy, probability judgment, surprisingness.

1 Introduction

Statisticians know that large samples are more likely than small samples to approximate the true population distribution of random events such as coin flips or die rolls. But people often ignore this distinction, and behave as though small samples are just as representative — what Tversky and Kahneman termed the "law of small numbers" (1971). Since people believe that small samples should be representative of underlying random distributions, they tend to think of streaks (and low-probability events more generally) as anomalies and suspect that a non-random process is interfering. Thus when streaks occur in random processes, as is likely to happen over time, individuals display a variety of biases that reflect their unease with the fact that the sample does not represent the distribution of the underlying population.

The most famous instantiation of this logic is the gambler's fallacy (Tune, 1964; Crites, 2003; Darke & Freedman, 1997; Gold, 1998) — the belief that a streak is more likely to end than chance would dictate. For example, after a coin flip yields three successive "heads," participants are more likely to bet that the next flip will be tails than after a mixed sequence that conforms to their schemata of what a random sequence will look like. This is a logical extension from the belief in the law of small numbers. If an individual expects a short sequence of coin flips to approximate the true distribution of all coin flips, then after a short streak of heads, tails must be more likely in future flips to "even out" the distribution. The end result is that by believing this, people act as though successive trials of a random event are not independent.

The gambler's fallacy is a bias in which people make inferences about future random events based on the outcome of previous events. However, the law of small numbers should also lead to inferences about unknown past events based upon knowledge of subsequent outcomes. We call this the *retrospective gambler's fallacy*: when making inferences about a series of random events that have occurred, people will show systematic biases in line with their conceptions of randomness.

The most straightforward instantiation of the retrospective gambler's fallacy would be formally identical to the gambler's fallacy, only in the past; after observing a coin land on three successive "heads," an individual might surmise that the flip immediately before the observation was a tails. While this would be a natural extension of the gambler's fallacy, it has, to our knowledge, never been tested. However, in this paper we explore other instantiations that are more novel and unique to the retrospective gambler's fallacy; namely, that unlikely outcomes will be perceived to have come from longer sequences than seemingly more common outcomes.

When individuals witness a low-probability event in a random series, what conclusions might they draw about

*This material is based on work supported under a National Science Foundation Graduate Research Fellowship. The authors thank Tom Griffiths, Neal Van Os, Pamela Sawyer, Ed Winiewski, Jessica Choplin, Art Markman, Katherine Kixmiller, Nick Davidenko, Bridgette Martin, Noel Brewer, Dave Budescue, Jonathan Baron, two anonymous reviewers, Dara Wathanapaisan, and Chris Olivola for advice and support. Address: Daniel M. Oppenheimer, Princeton University Dept of Psychology, Green Hall, Princeton, NJ 08544. Email: doppenhe@princeton.edu.

what preceded it? The law of small numbers suggests that people will not believe that rare events should occur in small samples of stochastic processes. Thus upon seeing a rare event, they must either abandon the notion that the process is stochastic, or abandon the notion that the sample is small. While there are documented situations in which participants reject the notion of randomness (e.g., the hot-hand fallacy, Gilovich, Vallone & Tversky, 1985), in this paper we focus on the other possibility: assuming the sample must have been large.

Real-world examples of this fallacy readily come to mind. If we hear that a teenager has unprotected sex and becomes pregnant on a given night, we might infer that she has been engaging in unprotected sex for longer than if we hear she had unprotected sex but didn't become pregnant, whereas, to our knowledge at least, the probability of becoming pregnant as a result of each intercourse is independent of the amount of prior intercourse. If you hear that two of your neighbours played the lottery yesterday and one of them won, our guess is that you would be more likely to assume that the one who has been playing for ten years has won rather than the one who just played this week for the first time in her life. Although it is true that a greater number of trials makes it more likely that the event will happen *at some point in the series*, the number of past trials has no relation to the outcome of the particular trial sampled, so witnessing an unlikely event in isolation says nothing about the number of prior trials.

Indeed, this bias may even speak to theories of the origin of the universe. Puzzling over the unlikelihood of the perfect alignment of all the variables enabling our universe to be stable and life to emerge (the "fine-tuning" problem), an influential school of thought has countered that as improbable as our universe is, it was inevitable for such a universe to exist if we assume that a great number of universes varying randomly on the important variables either co-exist at any given time or have succeeded each other for a very long time (see Misner, Thorne & Wheeler, 1973, cited in Hacking, 1987). For example, Leslie (1989, p. 70) writes that "the presence of vastly many universes very different in their characters might be our best explanation for why at least one universe has a life-permitting character." In other words, the "best explanation" for a low-probability event is that it is only one in a multitude of trials, which is the core intuition of the reverse gambler's fallacy. Yet others have criticized this reasoning by pointing out that however improbable our universe is, its occurrence tells us nothing about the existence of prior trials (Hacking, 1987), nor about the likelihood of co-existing multiple universes (White, 2000). The philosophical debate is focused on whether such arguments reflect the fallacy, but it provides little evidence as to whether people actually ever exhibit such a fallacy about more mundane matters. Note also that Hacking and White argue that the fallacy is demonstrated by believing in the very existence of prior trials, whereas we investigate a weaker form, where the existence of prior trials is a given, but the number of such trials is the focus.

We predict that in an unambiguously stochastic domain — such as a coin toss or a die throw — people will believe that the series of events has been occurring for longer after witnessing a seemingly unlikely event, than if no such event is observed, even if the event, sampled in isolation, in fact says nothing about the prior number of trials.

2 Study 1

2.1 Method

Participants and procedure. One hundred and eight Stanford University undergraduates filled out a survey as part of a course requirement. The survey was included in a packet of unrelated one-page questionnaires. Packets were distributed in class, and participants were given a week to complete the entire packet.

Stimuli and design. Two versions of a survey were created. Participants were asked to imagine that they walk into a room, and as they enter they observe a man flipping a coin five times. In the "streak" conditions, participants were told that all five flips came up heads. In the "non-streak" condition, participants were told that the coin landed on heads three times and on tails twice. Participants in both conditions were then asked to estimate, in an open-ended format, how many times the man had flipped the coin before they had entered the room.

2.2 Results and discussion

One outlier was eliminated for providing an estimate greater than five standard deviations from the mean. Participants believed that a sequence of coin flips was nearly twice as long before a streak (M = 16.2) than when there was no streak (M = 8.7), and this difference was significant, $t(106) = 2.17$, $p < .05$, Cohen's $d = .42$.

The results of Experiment 1 provide initial support for the existence of the RGF. In observing a series of independent events, one should learn nothing about the number of past events by witnessing subsequent events in the sequence. A critic might argue that if the goal of the coin-flipper was to achieve a streak of five heads, and the flipper kept going until reaching that goal, then the observation of the streak actually would be suggestive of a lengthier sequence. After all, the expected number of total flips necessary to observe a sequence of five heads is greater than the average expected number of total flips necessary to observe a sequence containing a mixture

https://doi.org/10.1017/S1930297500001170 Published online by Cambridge University Press

of heads and tails.[1] However, in this experiment it was clearly specified that the particular streak of five flips that was observed was determined by chance elements (when the participant happened to walk into the room) with no stated goal of achieving this particular outcome. Under these conditions, the objective expected total length of the sequence is unrelated to what is observed. Moreover, explicitly stating that the observation of that particular set of coin flips was arbitrary undermines the possibility of pragmatic inference that the flipper was actively trying to produce a streak. However, participants behaved as though the observation of a streak was evidence that the sequence was relatively long.

Study 1 focused on a particular category of rare event: a streak. It is unclear at this point whether retrospective reasoning about randomness is sensitive specifically to streaks or whether this phenomenon can be observed in response to any event that is perceived to be rare/unlikely. That is, like streaks, single events that are rare may not be expected to occur in small samples — the man who hits the slot machine jackpot might be perceived to have been playing for a longer time before winning, even though a jackpot is a one-time event. Therefore, the logic underlying the RGF for streaks should also apply to other rare single events. Study 2 was therefore designed to extend the findings beyond responses to streaks.

Additionally, Study 2 was designed to address several shortcomings in the first experiment. First, the scenario in the first study may have been ambiguous; while the number of each outcome (heads or tails) was given to participants, the order of those outcomes was not. In the streak condition, the order was unequivocal, as all five outcomes were heads. However, in the non-streak condition, the three heads and two tails could have been in a number of orders, ranging from alternating (HTHTH) to streaky (TTHHH). It could be that this ambiguity shook the confidence of participants and caused them to reduce their estimations of sequence length. This issue was addressed in Study 2.

Finally, it has long been known that people reason heuristically about subjective probability based on how representative a sequence is of an individual's notion of what a random series should look like (Kahneman & Tversky, 1972). In Study 1, the type of sequence that was objectively less likely was also less representative of randomness. Thus, the results of Study 1 do not shed light on whether the RGF is driven by representativeness or by the objective likelihood of a sequence.

Accordingly, Study 2 had four goals: to remove any confounds created by the ambiguity in the instructions from Study 1, to tease apart the influence of representativeness vs. objective likelihood in eliciting the phenomenon, to determine if the phenomenon could be observed in rare single events in addition to streaks (Study 2a), and to determine if the effect would obtain in a within-subject design (Study 2b).

3 Study 2a

3.1 Method

Participants and procedure. Eighty Stanford University undergraduates participated to fulfil part of a course requirement. The survey was included in a packet of unrelated one-page questionnaires. Packets were distributed in class, and participants were given a week to complete the entire packet.

Stimuli and design. Three versions of a survey were created, and each participant received only one version. Participants were asked to imagine that they are in a casino and happen to pass a man rolling dice. In one version of the survey, participants were told they witnessed three dice being rolled which all came up 6's. In a second condition, they witnessed three dice being rolled, two of which came up 6's and one of which came up a 3. A final condition told participants that they witnessed the rolling of two dice, both of which came up 6's. Although rolling two 6's on two dice is objectively twice as probable as rolling two 6's and a 3 on three dice, the latter sequence should be perceived as more representative of randomness (Griffiths & Tennenbaum, 2003). That is, it is more representatively random for the outcome of multiple die rolls to show several numbers than to have only a single number come up multiple times. Participants in all conditions were then asked to estimate, in an open-ended format, how many times the man had rolled the dice before they had entered the room.

3.2 Results

Participants believed that a sequence of die rolls was more than three times as long when a set of three 6's were observed (M = 34.2) than when there were only two 6's (M = 10.6), which in turn was believed to be longer than the representatively random sequence of two 6's and a 3 (M = 3.2).[2] The differences between groups was reliable, omnibus $F(2, 77) = 4.8$, $p < .05$, Cohen's $f = .18$. Pairwise

[1] To illustrate this, we created a simulation program in R (R Development Core Team, 2009). With 10,000 iterations, the median number of trials to get 5 heads in a row was 43 (mean = 63), while for only 3 it was 10 (mean = 14) (For a more systematic computational approach to "waiting times," see Hombas, 1997, cited in Nickereson, 2007.)

[2] As in Study 1, we scanned for outliers. However, in Study 2 there were no obvious outliers, so all data points were included in the analysis.

https://doi.org/10.1017/S1930297500001170 Published online by Cambridge University Press

comparisons showed that all differences between conditions were reliable as well, $t(47,48,57) = 1.94, 2.32, 2.65$, $p < .05$, Cohen's $d = .56, .67, .69$).

4 Study 2b

A critic might argue that in Study 2a participants had to pick a number, and thus could not indicate that they felt that there was no basis for judgment. Of course, were that the case one would expect them to answer randomly. The fact that there were reliable trends indicates that participants behaved as though they believed it possible to make inferences. That said, a stronger refutation to this criticism is to allow participants to indicate that they don't believe there is a difference between events prior to rare or common events. Thus, in Study 2b we ran a slightly altered version of Study 2a, but using a within-subject design.

4.1 Method

Participants, procedure, and design. Thirty-one participants were recruited from the Princeton University student center and compensated with candy. Participants were provided with the same basic scenario as in Study 2a in which they imagine observing a man rolling three dice. The man either rolled triple 6's in the "rare" condition or a 2, 4, 5 in the common condition. Participants were then asked to estimate, in an open-ended format, how many times the man had rolled the dice before they had entered the room. Participants responded to both conditions; the order of presentation was counterbalanced.

4.2 Results

Three participants were eliminated from the analysis for providing non-numerical answers (e.g., "many"). Participants believed that a sequence of die rolls was more than three times as long when a set of three 6's were observed ($M = 20.4$) than when a 2,4,5 was observed ($M = 12.7$). A paired-samples t-test run on log-transformed data confirmed that this difference was reliable, $t(27) = 2.67$, $p < .05$., Cohen's $d = .88$. It is worth noting that half of the participants provided the same answer for both scenarios. This suggests, that while the phenomenon is quite prevalent, there are individual differences in susceptibility. However, as half of the participants did show the bias, even in a conservative within-study design, it seems reasonable to presume that the effect is not merely an artifact of forcing participants to provide an answer.

4.3 Discussion of Studies 2a and 2b

Study 2 successfully replicated Study 1 using a different domain. This scenario involved a single rare event, rather than a streak in a series, suggesting that individuals do not expect rare events to occur in small samples, and that this expectation biases estimates of the number of total events that have occurred. Furthermore, the fact that individuals believed that two 6's came from a longer series than two 6's and a 3 is consistent with the predictions of the representativeness heuristic (Kahneman & Tversky, 1972). Moreover, the fact that the effect was strong enough to be observable in a conservative within-subject design (Study 2b) is testament to its robustness. As before, we stress that, although unlikely events are more likely to occur at some point in a longer sequence,[3] the number of prior trials is independent of the outcome of any particular trial.

At this point the demonstration of the RGF has been limited to two fairly artificial domains. This led not only to questions about the robustness of the effect but also about its generalizability. While the real world examples of teenage pregnancy and lotteries discussed in the introduction suggest that the RGF could apply to everyday events, Studies 1 and 2 examined only aleatory domains. This had the advantage of ensuring conditional independence and thus serving as a rigorous arena for an initial investigation. However ultimately this sort of reasoning should apply to a wider array of events, a possibility we explore in Study 3.

5 Study 3

5.1 Method

Participants. Thirty-one participants were recruited from a list of Princeton University students and staff who had signed up to be part of a paid subject pool. Participants were paid $8 for a half hour's worth of lab activities, which included the present study as well as several unrelated studies.

Stimuli and procedure. Sixteen stories were created which described an activity that a person was engaging in and the outcome of that activity. Because of a typographical error the results of one story were not included in the analysis[4]. Two versions of each story were constructed, one with a common outcome and one with an uncommon outcome (see Table 1 for brief versions). For example, "A little boy is playing in the sand at the beach and finds

[3] Again using an R simulation program with 10,000 iterations, the median number of throws was 150 to get a triple 6, 25 to get a double 6, and 51 to get two 6 and a 3.

[4] "A woman was stuck in traffic" was accidentally transformed into "a woman was struck in traffic."

https://doi.org/10.1017/S1930297500001170 Published online by Cambridge University Press

Table 1: Summaries of stories used in Study 3, and descriptive statistics broken down by rare vs. common ending. Note: TM is the 20% trimmed mean.

Story Version	Estimated number of prior trials				Likelihood ratings	
	Mean	SD	TM	Median	Mean	SD
A man eats undercooked meat and						
... gets sick	0.46	0.72	0.25	0	6.03	2.17
... does not get sick	26.40	76.68	3.67	3	4.40	2.25
A basketball team has a						
... five-game winning streak	6.92	6.18	5.67	4	4.24	2.05
... one-game winning streak	3.06	2.73	2.73	3	8.46	2.05
A man's train is						
... three hours late	16.67	27.95	8.75	7	2.07	1.60
... five minutes late	24.80	33.41	13.56	10	7.70	1.86
A man purchases a lotto ticket and						
... wins	78,287.46	276,955.33	857.89	150	1.30	0.79
... does not win	39.44	56.52	22.60	11	8.73	2.61
A man cheats on his taxes and						
... gets caught	10.77	16.62	5.56	2	4.30	2.26
... does not get caught	2.26	2.45	1.86	2	5.70	2.07
A man buys a bottle of soda and						
... wins a prize	637.08	1,696.04	141.25	150	1.57	0.73
... does not win a prize	67.35	241.06	3.64	1	9.2	1.37
A woman enters a raffle for concert tickets and						
... wins	12.69	26.40	5.67	5	3.07	1.89
... does not win	3.91	3.28	3.23	3	7.97	2.13
A telemarketer calls somebody and						
... makes a sale	146.92	261.52	82.78	75	2.53	1.17
... gets hung up on	115.25	237.99	59.20	50	6.97	2.16
A man pets a strangers dog and						
... the dog bites him	26.00	22.43	21.56	15	3.77	2.11
... the dog is friendly	21.31	51.11	7.60	5.5	6.83	2.09
A boy is playing on the beach and						
... finds a fish skeleton	11.42	14.56	7.00	8.50	3.17	1.93
... finds some shells	4.91	3.25	4.23	4	9.03	1.07
A woman is playing miniature golf and						
... gets a hole in one	5.75	4.16	4.75	5	3.70	1.49
... takes four strokes on the current hole	5.59	6.56	3.82	3	5.83	2.12
A man enters a restaurant and is seated						
... near a friend he hasn't seen recently	4.25	5.03	3.12	3	2.87	2.16
... near a window	8.09	9.49	5.95	5	5.63	1.52
A man goes fishing and						
... catches a very large fish	22.5	15.66	18.75	18.5	4.20	1.52
... catches only fish that are too small to keep	5.62	5.63	4.60	3	7.60	1.81
A man buys a guitar at a yard sale which						
... turns out to be worth a lot of money	7.77	5.26	7.22	10	2.60	1.65
... turns out to be out of tune	2.79	2.02	2.68	2	8.40	1.89
A woman has unprotected sex and						
... gets pregnant	7.31	7.65	6.11	3	5.97	1.99
... does not get pregnant	11.71	23.86	4.73	4	4.67	1.77

some shells" (common) or "... finds a fish skeleton)" (uncommon). Participants were randomly assigned to either the "common" or "rare" condition, and received stories of only one type. In an open-ended response prompt, participants were asked how many times the protagonist had engaged in the activity prior to the current story. For example "How many times has he been to the beach before?"

After answering all of the questions, participants were distracted for approximately ten minutes while they filled out unrelated surveys. They were then provided with the scenarios again, but this time with both outcomes, and asked to rate on a 10-point scale how likely each outcome was to have occurred, e.g. "A little boy is playing in the sand at the beach. On a scale of 1–10 how likely is he to find (a fish skeleton/some shells) (1 = *not at all likely*, 10 = *extremely likely*)?" The order of presentation was counterbalanced. This served as a manipulation check to ensure that the rare events were indeed perceived as more rare, and as a possible measure of mediation by perceived likelihood.

5.2 Results and discussion

The data from one participant were excluded for failure to follow the instructions. Additionally, responses given in non-numeric form (e.g., "many") were excluded (17 instances). Answers that were provided in a range were set to the mean of the range (e.g. "2–3" entered as "2.5") for the purpose of analysis (11 instances).

The remaining 435 observations were log-transformed to address the skewness of distributions [using log(x+.5) to deal with zeros], and analyzed as a mixed-effect model using lmer() from the lme4 package for R (Bates & Maechler, 2009). Specifically, we modelled random intercepts for the 15 remaining stories and the 30 remaining participants, treating version (rare or common) as a fixed factor (Model 1). As recommended by Baayen, Davidson, and Bates (2008), we used the pvals.fnc() function of languageR (Baayen, 2008) with 100,000 iterations to gauge significance. The results of these analyses are presented in Table 2. As predicted, we found that the contrast for version was highly significant, $t = 2.7$, $p_{MCMC} < .005$ — participants estimated a larger number of prior trials when they received the rare version than when they received a common version (see Table 1 for by-item means, medians, and 20% trimmed means — Wilcox, 2001).

Although rarity was manipulated between participants, all participants later rated how likely they thought both version of each story would be. However, to test mediation, we considered only ratings of likelihood corresponding to the version that participants encountered in their condition. First, versions designed to appear more surprising were on average rated less likely (Model 2), $t = -14.09$, $p_{MCMC} < .0001$. Second, when we included perceived likelihood for the relevant version of the story to the initial model (Model 3), we found that likelihood was a significant predictor, $t = -3.79$, $p_{MCMC} = .0003$, but that version was no longer a significant predictor, $t = .58$, $p_{MCMC} = .54$. Thus it looks like the difference in estimated number of trials observed between versions is mediated by the difference in average estimated likelihood.

Study 3 replicated the effects of Studies 1 and 2 in a variety of real-world domains. This demonstrates the robustness of the effect and implies that the RFG is not limited to situations when people believe an item is amenable to formal statistical analysis. Indeed the range of domains in which the RGF can be shown suggests that this sort of reasoning may be common outside of the laboratory.

6 General discussion

Three studies demonstrated that people believe that events perceived to be unlikely come from longer sequences than events that seem more probable, a consequence of reasoning retrospectively about random events as though those events were not independent. In Study 1, participants estimated a greater number of trials preceding a streak of coin tosses than a more typically random sequence. In Study 2, participants believed that more trials preceded seemingly unlikely die rolls than a more common-looking one, regardless of whether they considered such events in isolation or side by side. Finally, Study 3 showed that this reasoning generalizes and is prevalent in reasoning about rare and common events more broadly. Just like John Leslie writing about the fine-tuning problem in cosmology (1989), participants seem to think that the "best explanation" for an unlikely event is that it comes at the tail-end of many previous trials.

The retrospective gambler's fallacy may arise from the confusion between generating an unlikely event at some point in a sequence vs. at a particular point in the sequence. If you know that an unlikely event happened at some point in a sequence, all things being equal, it is rational to assume that the sequence was longer than one that did not contain the unlikely occurrence (see Footnotes 1 and 3). If two men had been put in a room and told to roll dice until getting a triple six, the first one to come out is indeed expected to have rolled more times. But if you just ask two men to roll dice continuously and you walk into the room at a random point in time, seeing that one of them just rolled a triple six doesn't tell you anything about how many times he rolled before you walked in. Even though the number of trials increases the probability that a given event will be obtained at some point in the sequence, it doesn't change the probability on any given trial.

https://doi.org/10.1017/S1930297500001170 Published online by Cambridge University Press

Table 2: Mixed-effect model equations for Study 3. Model 1 shows the effect of condition on estimates, while Models 2 and 3 serve to test mediation by likelihood ratings.

			s^2	b	SE	t	p_{MCMC}
Model 1 (Predicting log[estimates + .50])							
Random	Subject	Intercept	.27				
	Story	Intercept	1.18				
		Residual	3.86				
Fixed	Version (rare)			0.74	0.27	2.73	.006
		Intercept		1.23	0.33	3.71	<.0001
Model 2 (Predicting perceived likelihood)							
Random	Subject	Intercept	0.13				
	Story	Intercept	0.32				
		Residual	5.00				
Fixed	Version (rare)			−3.54	0.25	−14.09	<.0001
		Intercept		6.98	0.22	31.66	<.0001
Model 3 (Predicting log[estimates + .50])							
Random	Subject	Intercept	0.28				
	Story	Intercept	1.17				
		Residual	3.74				
Fixed	Version (rare)			0.18	0.31	0.58	.54
	Likelihood			−0.16	0.04	−3.79	.0003
		Intercept		2.35	0.45	5.27	<.0001

Note: Analyses conducted using lmer() function of the lme4 package for R. Markov Chain Monte Carlo p-values (p_{MCMC}) computed using pvals.fnc() function of the languageR package for R, with 100,000 iterations.

This may be an instance of what Kahneman and Frederick (2002) call attribute substitution — the notion that people presented with a difficult question will often substitute a different question that they can answer. In the current studies, the question of series length is difficult, as there is no "right" answer. However, participants did not answer randomly, as evidenced by the reliable differences between conditions. One explanation for these differences is that participants substituted the length of the particular series they were being asked about with the number of events necessary for them to expect such an unlikely event to occur. Nonetheless, while there is no "right" answer, we believe that making inferences about sequence length from single observations is both natural and common. This is evident in the fact that the findings generalized to such a wide array of domains in Study 3.

These studies raise an interesting puzzle about the law of small numbers. Original demonstrations of the law of small numbers (Tversky & Kahneman, 1971) suggested that people were insensitive to sample size when making judgments about random events. However, in our studies as the perceived rarity of an event increases, people estimate that it came from a larger sample, which suggests that people do indeed have an intuitive conception of the importance of sample size (even if they are mis-applying such conceptions in this instance). The literature on intuitive statistics is full of contradictory evidence about the extent to which people consider sample size in statistical reasoning (for a review, see Sedlmeier & Gigerenzer, 1997). Investigations into the RGF could lend new insight to this problem.

The fact that the magnitude of the bias appears to be related to participants' beliefs about the rarity of the outcome has important methodological implications. In-

https://doi.org/10.1017/S1930297500001170 Published online by Cambridge University Press

vestigations on beliefs about subjective probability estimates can be hampered by participants "knowing the right answer." For example, anybody who has taken a class in statistics should be able to tell you that there is a 50% chance of flipping a head even after five consecutive tosses of tails. However, when the task is subtle enough, even statistically sophisticated social scientists exhibit biases due to belief in the law of small numbers (e.g. Tversky & Kahneman, 1971). This suggests that, while individuals can be trained to give the normatively correct answer to a particular question, the underlying cognitive mechanisms for reasoning about randomness remain unchanged — it just takes more subtle and nuanced tasks to measure them.

Because many students have been exposed to the gambler's fallacy in psychology and statistics courses, it has become more difficult to measure reasoning about stochastic processes using traditional methods. The RGF offers a new tool in the repertoire for studying such topics. Not only is it a novel paradigm that participants are likely to be unfamiliar with, but by its very nature it prevents participants from computing values of how probable things are. Participants are never explicitly asked for probability estimates, so such computations would not give them the right answer.

In addition to their methodological implications, these findings have important theoretical implications. While this set of studies focused exclusively on judgment/inference about unobserved past events, there is reason to think that memory may similarly be biased by belief in the law of small numbers. There is a large literature demonstrating that memory is reconstructive (e.g,. Bartlett, 1932; Bower, Black, & Turner, 1979) and that memories are constructed so as to be consistent with prior beliefs (Abelson, 1981). Therefore, the beliefs that people hold about the nature of randomness could reasonably be expected to influence their memories for events that are perceived to be generated by stochastic processes. Indeed, recent research has shown that this can often be the case (Olivola & Oppenheimer, 2008). As people's schemas of randomness do not include the presence of long streaks, when people are exposed to sequences with long streaks they often remember the streaks to be shorter than they really were, damaging the accuracy of overall recall.

Aside from memory biases, there are also potential real world applications of the phenomenon. For example, the RGF may tie into people's belief in a just world (e.g. Lerner & Miller, 1978). People may be more inclined to "blame the victim" of unfortunate rare tragedies, on the premise that the victim must have been engaging in risky behaviors for a long time. Similarly, people may be willing to attribute rare and lucky successes to the notion that a person had been working towards that outcome for a long time and thus deserved it. In another vein, people who are biased in their estimates of how many trials preceded a perceived rare or common outcome might budget their time and resources inefficiently, akin to the planning fallacy (Buehler, Griffin & Ross, 1994).

The world is full of stochastic processes, and belief in the law of small numbers is a powerful principle for understanding reasoning about randomness. A thorough understanding of such reasoning processes requires that we not only examine how they influence our predictions of the future, but also our perceptions of the past.

References

Abelson, R. P. (1981). The psychological status of the Script concept. *American Psychologist, 36*, 715–729.

Baayen, R. H. (2008). languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics." R package version 0.953.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effect modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Bartlett, F. C. (1932). *Remembering — A study in experimental and social psychology*. Cambridge; Cambridge University Press.

Bates, D., & Maechler, M. (2009). lme4: Linear mixed-effects models using S4 classes. R package version 0.999375–31. http://CRAN.R-project.org/package=lme4

Bower, G. H., Black, J. B., and Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology 11*, 177–220.

Buehler, R.,Griffin, D., & Ross, L. (2004). Exploring the planning fallacy: Why people underestimate their task completion times. *Journal of Personality and Social Psychology, 67,* 366–381.

Crites, T. W. (2003). What are my chances? Using probability and number sense to educate teens about the mathematical risks of gambling In: H. J. Shaffer, M. N. Hall, J. Vander Bilt, E. George, & T. M. Cummings (Eds.) *Futures at stake: Youth, gambling, and society* (pp. 63–83). Reno, NV: University of Nevada Press.

Darke, P. R., & Freedman, J. L. (1997). Lucky events and beliefs in luck: Paradoxical effects on confidence and risk-taking. *Personality and Social Psychology Bulletin, 23,* 378–388.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology, 17*, 295–314.

https://doi.org/10.1017/S1930297500001170 Published online by Cambridge University Press

Gold, E. (1998). The gambler's fallacy. (Doctoral Dissertation: Carnegie Mellon University, 1998) *Dissertation Abstracts International,* 58, 7B.

Griffiths, T. L., & Tenenbaum, J. B. (2003). Probability, algorithmic complexity, and subjective randomness. *Proceedings of the 25th Annual Conference of the Cognitive Science Society.*

Hacking, I. (1987). The inverse gambler's fallacy: The argument from design. The anthropic principle applied to Wheeler universes. *Mind, 96*(383): 331–340.

Hombas, V. C. (1997). Waiting time and expected waiting time – Paradoxical situations. *American Statistician, 51*, 130–133.

Kahneman, D. & Frederick, S. 2002. Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, and D. Kahneman, Eds., *Heuristics and biases: The psychology of intuitive judgment*, pp. 49–81. New York. Cambridge University Press.

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

Lerner, M. J. & Miller, D. T. (1978). Just world research and the attribution process: Looking back and ahead. *Psychological Bulletin. 85,* 1030–1051.

Leslie, J. (1989). *Universes*. London: Routledge.

Misner, C. W., Thorne, K. S., & Wheeler, J. A. (1973). *Gravitation*, Foreman, San Francisco, CA.

Nickerson, R. S. (2007). Penney Ante: Counterintuitive probabilities in coin tossing. *The UMAP Journal, 28*, 503–532.

Olivola, C. & Oppenheimer, D. M. (2008). Randomness in Retrospect. *Psychonomic Bulletin and Review, 15,* 991–996

R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3–900051–07–0, URL `http://www.R-project.org`.

Sedlmeier, P. & Gigerenzer, G. (1997). Intuitions about sample size; The empirical law of large numbers. *Journal of Behavioral Decision Making, 10*, 33–51.

Tune, G. S. (1964). Response preferences: A review of some relevant literature. *Psychological Bulletin, 61*, 286–302.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 2*, 105–110.

White, R. (2000). Fine-tuning and multiple universes. *Noûs, 34*, 260–276.

Wilcox, R. R. (2001). *Fundamentals of modern statistical methods: Substantially improving power and accuracy.* New York: Springer-Verlag.