Global Research Unit
Working Paper #2018-014

Biased Beliefs About Random Samples: Evidence from Two Integrated Experiments

Daniel J. Benjamin, USC and NBER
Don A. Moore, UC Berkeley
Matthew Rabin, Harvard and Cambridge

Biased Beliefs About Random Samples: Evidence from Two Integrated Experiments

Daniel J. Benjamin
University of Southern California and NBER

Don A. Moore
University of California—Berkeley

Matthew Rabin[*]
Harvard University—Cambridge

Abstract: We report two incentivized experiments on four errors in reasoning about random samples: the Law of Small Numbers, Non-Belief in the Law of Large Numbers, exact representativeness, and "bin effects." We control for a variety of confounds that constrain prior work, test predictions of existing models, and assess the magnitudes of the biases. By asking each participant many different questions about the same data, we disentangle the biases from possible rational alternative interpretations. We find that no coherent model could jointly rationalize people's beliefs about random sequences with their beliefs about distributions of outcomes.

JEL Classification: B49

Keywords: Law of Small Numbers, Gambler's Fallacy, Non-Belief in the Law of Large Numbers, Big Data, Support Theory

[*] This paper supersedes the March 2013 paper titled "Misconceptions of Chance: Evidence from an Integrated Experiment" which reported on the findings of Experiment 1. We are grateful to Paul Heidhues, Ted O'Donoghue, and seminar audiences at Cornell, Harvard, Koc University, Penn, Stanford, UCL, LSE, WZB, and NYU, and conference participants at SDJM 2014 for helpful comments. We thank Yeon Sik Cho, Samantha Cunningham, Alina Dvorovenko, Tristan Gagnon-Bartsch, Kevin He, Han Jiang, Justin Kang, Tushar Kundu, Sean Lee, Hui Li, Stephen Yeh, Muxin Yu, and especially Lawrence Jin, Michael Luo, Alex Rees-Jones, Rebecca Royer, and Janos Zsiros for outstanding research assistance. We are grateful to Noam D. Elkies for providing math-professor responses. We thank NIA/NIH through grant T32-AG000186-23 to NBER and P30-AG012839 to the Center for the Economics and Demography and Aging at the University of California—Berkeley for financial support. E-mail: daniel.benjamin@gmail.com, dm@berkeley.edu, matthewrabin@fas.harvard.edu.

1. Introduction

Beliefs about random samples are a core input into risky decision making. This paper reports on two incentivized experiments that investigate four biases in reasoning about random samples. The first bias, which Tversky and Kahneman (1971) sardonically dubbed belief in the *Law of Small Numbers* (LSN), is the exaggerated belief that small samples will reflect the population from which they are drawn. In sequences of random events, LSN manifests itself as the *gambler's fallacy* (GF): People think that after a streak of heads, the next flip of a coin is more likely to be a tail. Rabin (2002) and Rabin and Vayanos (2010) formalize LSN, and argue that this bias can help explain some of the false (and costly) beliefs investors might have about stock-market returns. There is evidence that LSN influences decision making in a range of contexts, such as gambling (Croson and Sundali, 2005), including pari-mutuel lotteries (Suetens, Galbo-Jørgensen, Tyran, 2016; Terrell, 1994), and legal, lending, and pitching decisions (Chen, Moskowitz, and Shue, 2016).

The second bias is that people under-appreciate that large random samples will almost surely closely reflect the population. This bias leads people to overestimate the likelihood of extreme proportions in large samples, such as the probability of more than 950 heads in a thousand coin flips. Benjamin, Rabin, and Raymond (2016) build on evidence presented by Kahneman and Tversky (1972) to formally model this bias and explore its economic implications, dubbing it *Non-Belief in the Law of Large Numbers* (NBLLN).[1] Benartzi and Thaler (1999) provide evidence that NBLLN causes people to overestimate the likelihood of losing money in repeated gambles and long-term investing and thus contributes to overly conservative investing behavior.

The third bias is what Camerer (1987) called "*exact representativeness*" (ER): a tendency to overestimate the likelihood that even small samples will (nearly) exactly match the proportion of outcomes implied by the underlying probabilities.[2] Like LSN, ER is one way to make a specific prediction from Kahneman and Tversky's (1972) "representativeness heuristic," which they define

[1] NBLLN is pronounced with the same syllabic emphasis as "Ahmadinejad."

[2] Consistent with ER, Camerer (1987) and Grether (1980) found that their experimental participants overestimated the likelihood of a particular state of the world when the observed sample exactly matched that state's probabilities. Relatedly, Klos, Weber, and Weber (2005), while replicating Benartzi and Thaler's (1999) evidence for NBLLN, also found that experimental participants incorrectly believe the probability of a monetary outcome ending up within a fixed, small interval around the expected value increases with the number of repetitions of a gamble.

somewhat loosely as the belief that stochastic processes will produce samples very close to the underlying probabilities. To define ER distinctly from LSN, we mean by ER that people overestimate the likelihood that a sample of coin flips will be very close to 50% heads *even beyond* what would be implied by their belief in the gambler's fallacy. This bias sits in tension with NBLLN, and taken jointly, they suggest that people might tend to exaggerate both the likelihood that samples will be very close to expected proportions *and* very far, but underweight the likelihood of outcomes in between.

The fourth bias is more complicated, but is relevant in a broad array of contexts and central to interpreting our experimental results about the other three biases. This bias, which we call *bin effects,* is that people's beliefs are influenced by how the outcomes (about which participants are asked to report their beliefs) are categorized. In particular, the more finely a range of outcomes is divided when eliciting beliefs, the more total weight participants assign to the range.[3] Tversky and Koehler (1994) illustrate their closely related notion of "support theory" by showing that people assign greater total probability when asked separately about death by cancer, heart disease, and other natural causes (3 bins) than about death by all natural causes (1 bin). The implications of (and modeling challenges intrinsic to) such bin effects have been relatively little studied by economists; important exceptions are Ahn and Ergin (2011), who study such effects theoretically, and Sonnemann, Camerer, Fox, and Langer (2013), who explore how binning influences equilibrium prices in sports-betting markets and prediction markets. Our results provide additional evidence of this bias that might be of direct interest, but we focus on bin effects as an underappreciated confound in earlier evidence on biases such as NBLLN. Because the close-to-mean bins have higher probability than the far-from-mean bins in *all* related experiments we know, the inflated probabilities assigned to the far-from-mean bins could simply be the result of biasing all bins toward equal weight—per bin effects—rather than overweighting the likelihood of extreme sample proportions—per NBLLN.

[3] We use the term "bias" for all of the systematic deviations from normative rationality we explore. But it is worth noting that the errors have distinct psychological origins in ways that may render the term a poor fit. LSN and GF seem like biases in the sense that people have faulty models of random processes. NBLLN seems closer to ignorance of basic principles: people don't know the statistical principle, which is something that had to be discovered by very smart people with much effort. Bin effects (and possibly ER) are harder still to classify, and might merit the term "elicitation bias" as a sort of distortion in perceived or stated probabilities that seems tied to the questions people are being asked.

Our experiments are *integrated* in two senses, both of which enable us to draw conclusions about the biases jointly. First, our design is within-subject, with every participant answering every one of a large set of questions relating to each bias, and thereby provides rich identification of each participant's beliefs. Second, all questions are about the same, realized data sets. By asking different questions about the same data, our design lets us differentiate the biases from the possibility that participants did not understand or believe that the samples were truly generated from an i.i.d. process. Regardless of the process that generated the data, some of the biases we identify cannot co-exist as part of *any* set of coherent beliefs.

Our paper makes three main contributions. First, we provide evidence on the existence and magnitudes of the four biases after controlling for confounds that plague much of the earlier work. Second, we probe to what extent existing formal models of some of the biases accurately capture those biases. Third, we provide some of the only incentivized evidence on the biases we study (albeit with monetary incentives that are weak).[4,5] In addition to these specific contributions, we believe that our joint examination of the biases in our integrated experiment provides a clearer understanding of biased beliefs about random samples than has been possible from previous work, which to our knowledge always examined the biases separately.

Section 2 describes our experimental method. We studied participants' beliefs about sets of ten, one thousand, and (in Experiment 2) one million coin flips. In both experiments and for each sample size, we generated one million realizations of coin-flip data, and we elicited

[4] Despite large literatures related to the biases we study, there are surprisingly few incentivized experiments. Benjamin, Rabin, and Raymond (2016, Appendix D) find only 6 experiments (from 4 papers) on people's beliefs about sampling distributions (Peterson, DuCharme and Edwards, 1968; Wheeler and Beach, 1968; Kahneman and Tversky, 1972; Teigen, 1974). All of them are consistent with NBLLN, but none are incentivized, and only Kahneman and Tversky (1972) elicits beliefs about a sample size larger than ten. There is much evidence on LSN and the gambler's fallacy on beliefs about coin-flip sequences; see Oskarsson, Van Boven, McClelland, and Hastie (2009) and Rabin and Vayanos (2010) for reviews. However, this literature typically proceeds by showing that people think some equal-probability sequences are more likely than others without incentives or evidence as to how much more likely. An exception is Asparouhova, Hertzel, and Lemmon (2009), who document the gambler's fallacy in an incentivized laboratory experiment, but participants are not told the random process generating the binary outcomes. A number of (unincentivized) laboratory studies have found evidence of bin effects (e.g., Fox and Clemen, 2005). However, we are not aware of any work in economics or psychology focused on measuring people's beliefs that accounts for bin effects when interpreting the data—with the exception of Kahneman and Tversky (1972), who explicitly discussed in their paper the importance of holding constant the number of bins when comparing experimental participants' histogram beliefs across different sample sizes.

[5] In addition to incentives, we also made other design choices intended to add to the credibility of our findings. For example, all of our belief elicitations are framed in terms of frequencies of outcomes. As noted by Tversky and Kahneman (1983) and later researchers, posing problems in frequentist (as opposed to probabilistic) terms may mitigate some errors. We have not investigated whether these design choices mattered for our results.

participants' beliefs about the frequency of different sample realizations regarding this *fixed* set of realizations.

In Section 3, we describe our tests and findings for LSN/GF. We elicited participants' beliefs about the frequency with which streaks continued, for streaks of length 0 to 9. We also asked participants to make bets on which of two sequences of coin flips occurred more often. Their answers imply beliefs about how past—and future—flips affect the likelihood of heads. (Throughout, we refer to beliefs about "heads," but in fact all questions were randomized by question to ask about either heads or tails.) In Experiment 2, we also asked participants to bet on the likelihood of a given random flip anywhere in the full sample being heads given the outcomes of 1, 2, or 5 other random flips.

We find clear evidence for GF in both experiments. Most straightforwardly, across experiments and different settings, participants on average assessed a 44% to 50% chance that a first flip would be a head, but 32% to 37% chance that a flip following 9 heads would be a head. Analysis of median answers and individual variation suggest fewer than 50% of participants believed in GF, but it was far more common than bias in the other direction.[6] The data on betting which of two sequences is more likely is noisier, and the evidence for GF in that data is suggestive and not statistically significant. Consistent with our conjecture that motivated some of our betting questions, but contrary to existing formal models of LSN, we find suggestive evidence against what Oppenheimer and Monin (2009) called the "backward-looking" GF: We could discern no effect of knowledge of later flip outcomes on participants' predictions regarding earlier flips in the sequence. Also contrary to existing formal models of LSN, we find evidence of "long-distance GF" in Experiment 2: The mean prediction of the probability that a randomly chosen flip in the thousand- and million-flip sets would be heads given that 5 other randomly designated flips were heads is 40%—roughly the same likelihood participants believed for a flip following 5 heads in a row. These findings together suggest that the psychology of LSN manifests itself as the GF when people predict outcomes based on past sequences of flips, but it does not have the tight logical structure of negative serial autocorrelation that current models embed. In a setting of forward-

[6] For all analyses, we conduct both aggregate and individual-level results. We note in the text when the individual-level evidence might affect conclusions, and we otherwise relegate those results to Appendix G.

looking and non-long-distance GF, we estimate the parameters of Rabin and Vayanos (2010)'s formal model of LSN.[7]

Section 4 investigates bin effects. We elicited participants' histograms of beliefs about the frequency distributions of heads and tails in samples of size ten, a thousand, and a million, with the possible outcomes binned in several different ways. For example, for a sample size of ten, our 11-bin treatment elicited beliefs about the likelihood of each possible outcome from 0 heads to 10 heads, while our 2-bin treatment asked only about the likelihood of 5 heads (as opposed to anything else). The bin effects we found were stronger than we anticipated and, we suspect, stronger than most researchers might have expected—especially given that our questions were repetitive within-subject elicitations about objective probabilities. Going from 11- to 5- to 3- to 2-bin treatments, the mean probabilities assigned to 5 out of 10 heads were 20%, 28%, 36%, and 39% in Experiment 1 and 28%, 32%, 33%, and 34% in Experiment 2.

Given bin effects, the question of how one elicits "true" beliefs is difficult (and perhaps ill-conceived). However, if we postulate the existence of what we call *root beliefs*—which are then operated on by bin effects when people report their beliefs—then under some assumptions and with the right kind of data, we can identify how the root beliefs are biased relative to true probabilities. Building on existing models of support theory (Tversky and Koehler, 1994; Ahn and Ergin, 2011), we formalize this approach and use it to try to disentangle biased (root) beliefs about frequency distributions of heads and tails from bin effects. The simplest implication of the assumptions we outline is that when beliefs are elicited with some categorization of outcomes including events A and B such that the true probabilities are $\pi(\mathrm{A}) \geq \pi(\mathrm{B})$ but participants' elicited beliefs are $P(\mathrm{A}) < P(\mathrm{B})$, then there is strong evidence of a bias in root beliefs (i.e., net of the bin effects). For instance, if observed beliefs are that (say) people believe that exactly 5 heads is more likely than 0-4 heads when asked their beliefs about 0-4, 5, and 6-10 heads, whereas we know that it is less likely, we infer that people are exaggerating the probability of 5 heads relative to 0-4.

In Section 5, we use this approach to study participants' beliefs about the frequency distributions of heads and tails in samples of size ten, a thousand, and a million. We test for overweighting of the extremes of the distribution as entailed by NBLLN and overweighting of the middle of the distribution as entailed by ER. The strength of the bin effects we find lends support

[7] While we know of no firm way to characterize the size of the effect in comparison to previous studies, our parameter estimates are smaller than those of Rabin and Vayanos's (2010) based on hypothetical questions in prior studies.

to our worry that some of the putative evidence of NBLLN observed previously may be misleading. For example, although all previous data we know of suggest that people overweight the likelihood of extreme outcomes in sample sizes of ten, our results show that this can be reversed when the bins are changed. In fact, taking bin effects into account, our data suggest that—contrary to earlier interpretations of the evidence (e.g., Benjamin, Rabin, and Raymond, 2016)—people have reasonably well calibrated frequency-distribution beliefs for samples of size ten.

For sample sizes of a thousand and a million, however, the result is that beliefs across different binning arrangements produce what appears to be a "real" NBLLN. For example, we find that when we elicited beliefs for 5 bins that have as close to equal likelihood as we could get—[0-487], [488-496], [497-503], [504-512], and [513-1000], with probabilities 21.5%, 20%, 18%, 20%, and 21.5%, participants' mean reported beliefs were 26%, 12.5%, 21%, 13.5%, and 27%, respectively. Since the observed beliefs for the extremes of the distributions are too large, and since bin effects would compress the beliefs *toward* equal likelihood, we believe our results provide strong evidence that people's root beliefs overestimate the likelihood of the extremes. Our results additionally indicate that people overestimate the likelihood of the *middle* bin. Because the magnitude of the middle-bin overestimation is greater than can be explained solely by the GF observed in participants' responses to other questions, it points to ER as a distinct bias. We interpret the results for the sample size of 1 million as similar, but the evidence is less consistent.

In Section 6, exploiting the integrated nature of our experimental design, we report tests that show that a coherent-but-mistaken model of the process generating the coin flips cannot reconcile participants' beliefs about sequences of coin flips with their beliefs about the distribution of outcomes. In both experiments, consistent with GF, participants report that following a streak of 9 heads, another head is roughly half as frequent as a tail. Since the nine other ways to get the outcome "9 heads out of 10" are surely believed to be at least as frequent as HHHHHHHHHT, the outcome "9 heads out of 10" is implied to be at least 20 times more frequent than the outcome "10 heads out of 10"—yet the fact that participants' histogram beliefs (about the same set of realizations) are reasonably well calibrated implies that they believe it is only 10 times more frequent. To conduct further tests, we prove a proposition about beliefs under the most plausible coherent model participants might hold: parameter uncertainty about the bias of a negatively autocorrelated coin. We show that, given a sample of coin flips generated by this model, if M randomly drawn flips from the sample are all heads, then the likelihood that the M+1st randomly drawn flip from the same sample is a head should be increasing in *M*. Our implementation of this test instead finds that people believe the likelihood is *decreasing* in *M*. Because participants' answers are internally inconsistent, our results cannot be fully explained by participants disbelieving or misunderstanding the experimental instructions or thinking that the coin flips were generated from a non-i.i.d. random process.

Section 7 reports robustness and additional analyses on survey-question order effects, participants' use of calculation tools, the financial costs of biases in the experiment, participants' effort, and the within-subject correlation across biases.

Section 8 concludes with a brief discussion of some of the broader implications of our results for economic theory, and for experiments and surveys intended to elicit beliefs. Whereas virtually all economic models of belief biases are "quasi-Bayesian"—assuming that people have an internally consistent but false model of the data-generating process (e.g., Barberis, Shleifer, and Vishny, 1998; Rabin, 2002)—our results imply that such models will not be able to predict some important aspects of people's beliefs. Moreover, the divergence we find between beliefs about sequences and beliefs about the distribution of outcomes implies that people's beliefs will depend on how they conceptualize the problem they face—which in turn implies that learning about how people frame such problems to themselves is a priority for future experimentation and theorizing. The strong bin effects we observe suggest they are potentially a major confound in efforts to measure beliefs via surveys.

Appendices A-D contain the design, analysis plan, and screenshots for the experiments. Appendix E contains proofs of our theoretical results. Appendices F-I contains discussion of additional analyses, our planned analyses annotated with the complete set of results, and appendix tables and figures. For reference, Appendix J reports the numbers underlying all figures.

2. Experimental Design

In February, 2010, we circulated among ourselves an experimental design and analysis plan document, which we finalized in March 2010 (see Appendix A). We conducted an experiment with 104 customers at a food court in Pittsburgh, PA, and circulated a paper based on this experiment (http://learnmoore.org/Barney/Barney.pdf). Based on the results from this experiment and new ideas for implementation, we finalized our Experiment 2 design and planned analysis in April 2014, and we preregistered it on the Open Science Framework (posted at https://osf.io/mjwbi/ and in Appendix B). We implemented Experiment 2 with a sample of 308 students at an experimental lab on the Berkeley campus.[8]

The flow of each experiment was similar: General instructions and explanation of incentives, a comprehension quiz, a practice histogram question, blocks of questions about beliefs that were the main body of the experiment, a math quiz, demographic questions, and finally questions about behavior during the experiment. Relative to Experiment 1, Experiment 2 expanded our sample size, sharpened the incentives, and (most importantly) added new questions that we realized after analyzing the data from Experiment 1 would help us more cleanly identify the belief biases.

Table 1 summarizes the designs of Experiments 1 and 2. The complete experimental materials for both experiments are available online (https://osf.io/mjwbi/). Screenshots from the experiments are in Appendices C and D. We describe the motivation for and details of the 54 questions in Experiment 1 and 88 questions in Experiment 2, together with the results, in Sections 3-5.

2.1. Experimental Samples

For Experiment 1, we recruited participants from a busy food court in downtown Pittsburgh, Pennsylvania. We aimed for 100 participants, and ended up with complete data from 104. Upon agreeing to participate, participants were taken to a room adjacent to the food court and answered questions on computers set up at cubicles. The median time to complete the experiment was 27 minutes. Among the 93-96% of participants who answered the demographic questions,

[8] The main text of the paper reflects all primary *ex ante* (before the 2010 experiment was run) and interim (before the 2014 experiment) hypotheses and planned tests that motivated the experiments; we relegate secondary analyses to Appendix F. However, our discussion often reflects sharper conceptualizations of the issues than we had when we ran the experiment, and some of our statistical tests were formulated after looking at the data. Specifically, we formulated the theory-based estimation of the parameters of Rabin and Vayanos's (2010) model in analyzing the results from Experiment 1 after seeing weaker-than-anticipated results, and some of our specific comparisons of bin effects across questions were unplanned. Moreover, when revising the paper to include Experiment 2, we re-labeled bin effects as a belief bias (rather than distinguishing it as what we called an "elicitation bias"). We also re-framed exact representativeness to be one of our biases of primary interest, in line with our original analysis plan, because Experiment 2 allowed us to more clearly distinguish it from the LSN. Furthermore, for both experiments we formulated our planned analyses of ER as if it were fully opposite to NBLLN, and we only realized when analyzing the results of Experiment 2 that we could test for their joint presence.

57% were male, with ages ranging from 18 to 69 with a mean of 34 years, and 71% self-reported income falling in the category "below $50,000."

In Experiment 2, participants were undergraduate students at the University of California at Berkeley and part of the subject pool of the Experimental Lab for Behavioral Research. We aimed for 300 participants, and ended up with complete data from 308. The median completion time was 40 minutes. When subjects arrived at the experimental laboratory, they were seated individually at computers. Among the 96-98% of participants who answered the demographic questions, 65% were female with an average age of 20. For details on demographics for both experiments, see Appendix Table H.1.

2.3. Instructions

In the general instructions to the experiments, we explained that we simulated samples of coin flips. For example, in Experiment 2, the instructions stated:

> There will be three groupings of questions on this survey:
>
> Ten-flip sets. For one group of questions, we began by having a computer simulate flipping a coin ten times. That made one set. Then we did it again and again and again…until we had a million ten-flip sets.
>
> Thousand-flip sets. For another group of questions, we did the same thing except that each set had a thousand flips in it.
>
> Million-flip sets. For the third group of questions, each set had a million flips in it (yes, a million sets of a million coin flips each—we really like flipping coins).
>
> We will randomly determine the order in which you encounter these three question groupings. Each simulated coin flip had equal chances of coming up heads and tails (i.e., it was a fair coin).

For both experiments, we generated these sets of 1 million samples using the pseudorandom number generator in Matlab.

In both experiments, we also included an instruction intended to discourage participants from trying to use the questions we asked to infer either the correct or the sought-after answers:

> For some of the questions, we will ask you to make judgments using numbers or ranges we provide. In some of these questions, we have chosen the examples and

> numbers literally randomly. At other times, we have carefully selected the examples to get your judgments in a wide range of scenarios. In fact, you will note that we often ask very similar-sounding questions; these questions may have similar answers, or very different ones. In all of these cases, the specific numbers in the question are NOT meant to indicate more or less likely answers.

Experiment 1's instructions did not mention calculators or other calculation tools, but participants in Experiment 2 (undergraduates in a computer lab) had access to such tools. We instructed them:

> You're welcome to take notes, do calculations on paper, or use a calculator if you like. It may or may not be useful to help you answer more accurately.

We asked participants at the end of Experiment 2 about their use of calculation tools. In Section 7 we report details, but we found little difference in results between those who reported using tools and those who did not.

2.4. Randomizations

To neutralize any bias toward believing heads or tails is more likely, we randomized any survey question for which it was relevant as to whether we asked about heads or tails. We explain other question-specific randomizations (including randomizing the numbers in the questions to try to neutralize inferences that participants might otherwise draw) in Sections 3-5 below, when we describe the belief questions we asked participants. We also randomized the order of question blocks and the order of questions within the blocks. These enable us to test for fatigue effects and for effects of inferences from questions asked earlier in the experiment on responses to later questions; see Section 7.

2.5. Payments and Incentives

We paid all participants both a base payment and payment contingent on their answers. The base payments for completing the experiments were a choice of either $3 cash or a $5 gift certificate valid at one of the food-court vendors in Experiment 1 and $5 in Experiment 2.

Participants accumulated "lottery tickets" throughout the experiment based on their answers.[9] The lottery ticket scheme differed across the two experiments. In Experiment 1, the probability of a participant winning a $50 additional prize was set equal to (#lottery tickets accumulated)/40,000. The mean number of lottery tickets accumulated was 3,370, yielding a mean probability of winning the prize of about 8%. Winners of this $50 prize were paid by check sent in the mail.

In Experiment 2, in order to sharpen the incentives for accuracy, we adapted a procedure proposed by McKelvey and Page (1990). We told participants that *each* of the questions on the survey had a prize associated with it. The probability of winning any given prize depended on the number of prize tickets earned for that question, where each ticket increased the chances of winning the prize by 0.05 percentage points. The maximum possible number of tickets one could earn for a given question was 100, so that the maximum chance of winning any given prize was 5 percent. Each prize was worth N^2 dollars, where N was the total number of prizes won by that participant.[10] Despite this modification, as we explain in Section 7 the incentives were in fact not very sharp in many of our elicitations.

3. The Gambler's Fallacy

In this section we report results testing the extent and nature of participants' belief in the gambler's fallacy and the law of small numbers. We report the most basic results in Section 3.1, structurally estimate the magnitude of these effects in Section 3.2, explore other ways of eliciting beliefs about sequences that speak to some of the anticipated and unanticipated limits to current models of GF in Sections 3.3 and 3.4, and summarize the array of results in Section 3.5. Our sample size and repeated-measures design afford fairly powerful statistical tests, and the key results we claim are

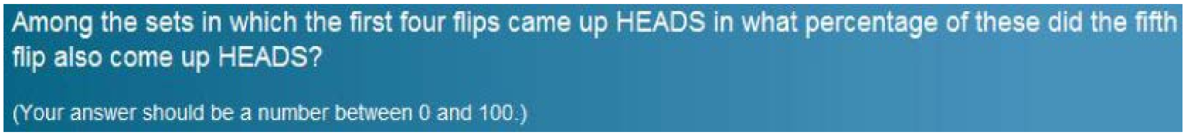
[9] Incentivizing participants with lottery tickets is a standard technique designed to induce risk-neutrality because, as argued by Roth and Malouf (1979), theoretically an expected-utility-maximizing participant should be risk-neutral over lottery tickets, in principle allowing the design of incentive-compatible revelation of beliefs.

[10] The instructions explained, "In the extremely unlikely event that you win prizes on all 66 questions, you will earn $4356 (66 prizes, each worth $66). We're not kidding. The reason we're paying you this way is that it gives you a strong incentive to do your best on every question: Winning a prize on any given question will not only earn you money on that question, but will also make it more valuable to win a prize on every other question."

highly significant. To avoid cluttering the text, we generally omit reporting *p*-values, presenting them only when they fall short of statistical significance at the (more stringent than usual) 0.005 level.

3.1. Evidence from Beliefs About Streak Continuation

The most straightforward evidence of the GF comes from eliciting the conditional probability of a head following specified streaks of heads. In Experiment 1, we tested for the presence and magnitude of GF solely in samples of size 10. One way we did so was to elicit beliefs about the frequencies with which streaks will continue. For example, we asked:



Among the sets in which the first four flips came up HEADS in what percentage of these did the fifth flip also come up HEADS?

(Your answer should be a number between 0 and 100.)

Participants were asked the probability of a head, conditional on all prior flips being heads, for each of the 9 streaks from 1 to 9 heads, on different screens and in random order. We also asked all participants a 10th question, at a random point, about the first flip: "What proportion of sets have a HEAD as the first flip?"

The resulting data, participants' mean and median beliefs about the frequency of a head on the *m*th flip given that the first *m*-1 flips were all heads, are displayed in Figure 1. Consistent with GF, the mean beliefs are decreasing as *m* increases from 2 to 10, dropping nearly monotonically from 46% after 1 head to the (statistically significantly) lower 32% after 9 heads.[11] We were worried, however, by the fact that participants' mean judgment of the probability of a head on the *first* flip was 44%, significantly below the obvious answer of 50%.[12] In trying to understand this puzzling result, we found that many participants reported a frequency of 10% for some values of *m*, suggesting that these participants were confused about the question. Figure 2 shows the data for the remaining 69 participants after excluding all participants who answered 10% for any value

[11] Hereafter, unless stated otherwise all statistical comparisons are from two-tailed *t*-tests when comparing mean beliefs and from Wilcoxon signed rank sum tests when comparing median beliefs. These are one-sample tests when comparing participants' beliefs to a specific value (such as 50%) and paired tests when comparing participants' beliefs across two situations (such as after 1 head vs. after 9 heads).

[12] Recall that these and other reports of "heads" were really balanced both heads and tails, so this deviation from 50% cannot be attributed to a bias in favor of expecting heads or tails.

of *m*. In this sample, the mean belief of 48% on the first flip is not statistically distinguishable from 50% ($p = 0.44$). As in the unadjusted data above, however, participants' mean beliefs are consistent with GF for $m \geq 2$, dropping almost monotonically from 48% after 1 head to 37% after 9 heads. But note that in both samples, contrary to existing models of GF, participants' mean belief following a single head was also indistinguishable from (and slightly more than) their mean belief about the first flip.

Relative to Experiment 1, we modified the procedure in Experiment 2 in two ways. First, we elicited beliefs regarding all three sample sizes: ten, a thousand, and a million. Second, both to hold constant the wording when adding these questions about other samples and to neutralize one possible misunderstanding that might have led to the anomalous "10%" responses, we changed the wording of the question as follows: "Some of the sets of 10 [1000 / 1 million] flips had Heads come up on the first [*j*] flips. For all of these, please estimate the percentage that also came up Heads on the next flip."

Figure 3 shows the evidence from Experiment 2. For all three sample sizes, the results replicate the same basic GF pattern as in Experiment 1: For $m \geq 2$, participants' mean belief about the likelihood of a head is smaller the longer the streak of prior heads. For sample sizes of ten, a thousand, and a million, the mean judged probability of a head after 9 heads is 35%, 37%, and 38%, in all three cases significantly smaller than the probabilities of 45%, 44%, and 44% after 1 head. Participants' mean beliefs appear virtually identical across the sample sizes, with the one exception being the probability of a head on the first flip (for which the question was: "Please estimate the percentage of 10-flip [1000-flip / 1,000,000-flip] sets that had heads come up on **Flip 1**"). For the sample size of 10, unlike in Experiment 1, we find that the mean probability predicted for head on the first flip is 50%. The mean probability of a head after 1 head is 45% (statistically different than 50%), and it declines monotonically to 35% after 9 heads. While the mean of 50% for the initial flip indicates we eliminated the confusion from Experiment 1, in fact once again participants predicted only 44-45% for even the first flip in the two other cases, indicating new confusion in the questions for those sample sizes.[13] For beliefs following 1 to 9 heads, the decline

[13] But the confusion was clearly different, as indicated by both the change in wording we employed and the fact that very few people predict 10%. There is evidence that the confused responses in the 1000-flip (but not 1-million-flip) questions may be at least partly due to misunderstanding of the conditional probability question: The mean reported unconditional probability of a head is statistically significantly lower if the unconditional probability was elicited

is nearly monotonic, but less rapid than in Experiment 1, suggesting perhaps that the Berkeley students in Experiment 2 suffered less from the GF than the shoppers in Experiment 1.

In contrast to the participants' mean belief, median beliefs show no sign of GF, virtually always equal to 50% in Experiment 1 (Figure 2) and always 50% for every sample size and every *m* in Experiment 2 (Appendix Figures I.1-I.3). Individual-level beliefs reveal that while many participants report the correct belief of 50%, GF is the predominant bias—much more common than the opposite biased belief that streaks continue or any other interpretable error. We confirmed this in several ways, including individual-level regressions. Some summary statistics on these streak questions make the finding clear, however: 28% of the 69 participants who never said "10%" in Experiment 1 and 44% of all 308 participants in Experiment 2 correctly said 50% for all their answers. Of the others, 26% of 50 in Experiment 1 and 66% of 171 in Experiment 2 always said answers weakly less than 50%. Only 1 participant in Experiment 1 and none in Experiment 2 always said weakly greater than 50%.

3.2. Structural Estimation of GF

There is a large previous literature looking at belief in GF in symmetric binary processes, often framed as coin flips. Different methods—production tasks, where participants write out what look to them like random sequences; recognition tasks, where participants select which of two sequences look to them more likely to come from coin flips; and prediction tasks, where participants are asked to predict next flip in a sequence—all indicate belief in the gambler's fallacy. So far as we know, none of these experiments are incentivized, and—more importantly—it is hard to assess the magnitude of any bias since all sequences are in fact equally likely. But the experiments nearly universally support belief in GF, with participants clearly believing randomness involves fewer streaks than it does. Although few direct comparisons are available,

before any of the conditional probabilities (37% if asked first vs. 44% if not asked first; $p = 0.02$); see Appendix G.I.A.3. The deviations from 50% for sample sizes of 1000 and 1 million are *not* due to a disproportionate fraction of participants reporting 10%, as seemed to be driving analogous behavior in Experiment 1, but rather to a variety of responses less than 50%, often by participants who correctly answered 50% to the 10-flip question (Appendix G.I.B.2.a. and G.I.C.2.a.).

and even these are confounded by methodological differences, our sense is the magnitude of GF that we observe appears to be weaker than has been reported in prior work.[14]

To facilitate comparisons of the strength of GF across settings, we use our streak data to estimate the parameters α and δ of Rabin and Vayanos' (2010) model[15]:

$$q_t = \omega - \alpha \sum_{k=0}^{\infty} \delta^{k+1} y_{t-1-k}, \tag{1}$$

where q_t represents the agent's belief regarding the t^{th} flip; y_{t-1-k} represents the outcome of the $(t-1-k)^{\text{th}}$ flip; ω parameterizes the bias of the coin; $\alpha \in [0,1]$ parameterizes the magnitude of the GF effect on the next flip; and $\delta \in [0,1]$ parameterizes the rate of decay of the GF effect on subsequent flips. The outcome variable y_{t-1-k} is equal to +1 if the $(t-1-k)^{\text{th}}$ flip was a head and -1 if it was a tail. The agent's perceived probability that the t^{th} flip will be a head, $p_t = (q_t + 1)/2$, is a rescaling of the belief variable $q_t \in [-1,1]$. We conduct the estimation using non-linear least squares on the first-difference of equation (1): $q_t - q_{t+1} = \alpha\delta^t$ (where every y_{t-1-k} equals either 1 or -1 because in the sequences we presented to participants, the realized outcomes were either all heads or all tails). For Experiment 1, the data we use are the mean judgments p_t shown in Figure 2 (all for sample size *N* = 10), and for Experiment 2, the mean judgments p_t in Figure 3 separately by sample size *N* = 10, 1000, and 1 million.[16]

The results are shown in Table 2. In all cases, we find positive point estimates for α, consistent with GF, and ranging from 0.02 to 0.16 (although the 95% confidence interval includes

[14] For example, Rapoport and Budescu (1997) find that after three consecutive heads, participants' belief that the next flip will be heads is only 30%, whereas we find that it takes nine consecutive heads to reach a comparably low level of belief (32% in Experiment 1 and 35% in Experiment 2).

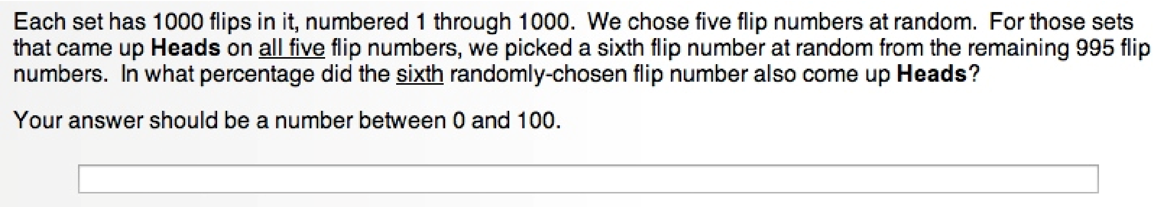
[15] Our equation differs from the corresponding equation (4) in Rabin and Vayanos (2010, p.736) in three ways. First, to more clearly distinguish notationally between an agent's perceived probability of a head and the past outcome of a coin flip, we change Rabin and Vayanos's notation of ε_t and ε_{t-1-k} to q_t and y_{t-1-k}. Second, Rabin and Vayanos allow the bias of the coin to be time-varying and therefore they subscript ω by *t*, whereas we do not. Finally, we correct a typo by writing δ^{k+1} (rather than δ^k).

[16] We report the parameter values estimated from the data on mean participants' beliefs in order to make transparent the relationship between the estimation and the discussion of the mean beliefs shown in the figures. For the data from both experiments, we have also estimated this regression on the individual-level data, imposing the same parameter values for all individuals but with standard errors clustered by individual. In all cases, the results are nearly identical, up to the third or fourth decimal place; see Appendix Table H.2.

zero in Experiment 1 and in the N = 1000 data in Experiment 2). The estimates for δ in our data range from close to one (indicating little decay of the GF) to roughly 0.6.[17]

3.3. Evidence on Beliefs about Flips From Randomly Chosen Locations

In Experiment 2 (but not Experiment 1), for each of the three sample sizes, we included three questions that asked about a random flip, conditional on the outcomes of 1, 2, or 5 others. For example:



Each set has 1000 flips in it, numbered 1 through 1000. We chose five flip numbers at random. For those sets that came up **Heads** on all five flip numbers, we picked a sixth flip number at random from the remaining 995 flip numbers. In what percentage did the sixth randomly-chosen flip number also come up **Heads**?

Your answer should be a number between 0 and 100.

The primary reason we asked these questions was to test whether participants' beliefs about the distribution of outcomes might derive from uncertainty about the bias of the coins; finding no positive correlation in such questions would let us reject the "parameter-uncertainty hypothesis," as we discuss below in Section 6. The unexpected finding of *negative* correlation indeed lets us reject this rationalization of NBLLN, but also provides evidence for a stronger (and harder to model!) form of GF.

Figure 4 shows the results. The results are not just qualitatively similar to those from the streak results shown in Figure 3, but even quantitatively similar. For the sample size of 10, the mean reported probability of a head after a head in 1, 2, and 5 randomly chosen locations is 46%, 41%, and 37%, quite close to (and not statistically different from) the probabilities after streaks of length 1, 2, and 5 of 45%, 41%, and 38%.[18] Just as in the streak data, but more problematically in this context, the mean beliefs for sample sizes of a thousand and a million are nearly identical to those for ten. These results on mean beliefs provide further evidence of the psychology of LSN: People expect to see balance in any small sample.

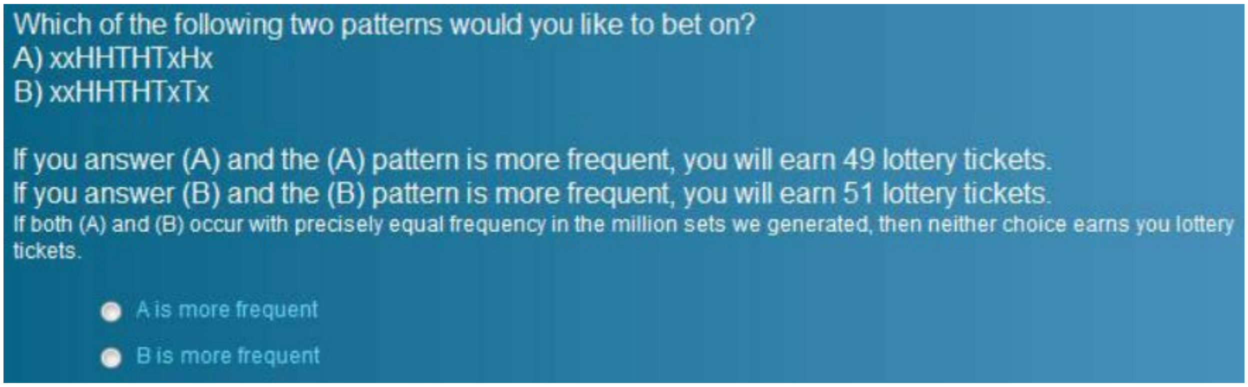
[17] For comparison, Rabin and Vayanos (2010) estimate that $\alpha \approx 0.2$ and $\delta \approx 0.7$ in Rapoport and Budescu's (1997) data. Our smaller α estimates imply that the GF is weaker in our data than in Rapoport and Budescu's (1997).

[18] As in the streak data, while mean beliefs show the GF, median beliefs are virtually always 50% (Appendix Figures I.4–I.6). The GF is the dominant direction of bias at the individual level: 43% of participants gave equal weights to 1, 2, and 5 randomly chosen locations, 32% of participants gave non-equal and weakly decreasing probabilities going from 1 to 2 to 5 random locations, and 8% of participants gave non-equal and weakly increasing responses.

We did not anticipate this finding, which we believe is important. The results in the thousand and million sample-size cases are inconsistent with existing models of GF, which predict belief in negative autocorrelation locally but not over long ranges. In fact, the belief that any subset of flips in the sequence will display LSN is not only inconsistent with *existing* quasi-Bayesian models, but with *any possible* quasi-Bayesian model: there is no logically consistent beliefs participants could have that would generate such beliefs.

3.4. Evidence on Bets about Sequences

The final source of evidence about GF comes from participants' bets. In Experiment 1, we asked ten questions in which participants bet on which of two sequences are more common among the one million samples of size ten. The instructions informed participants that the sequences were generated randomly. This randomization allows us to examine GF outside the context of extreme samples, such as streaks of all heads, and also makes it less likely that participants would draw an inference from our question regarding what answer we expect or think they "should" pick. The following is an example betting question[19]:



Which of the following two patterns would you like to bet on?
A) xxHHTHTxHx
B) xxHHTHTxTx

If you answer (A) and the (A) pattern is more frequent, you will earn 49 lottery tickets.
If you answer (B) and the (B) pattern is more frequent, you will earn 51 lottery tickets.
If both (A) and (B) occur with precisely equal frequency in the million sets we generated, then neither choice earns you lottery tickets.

○ A is more frequent

○ B is more frequent

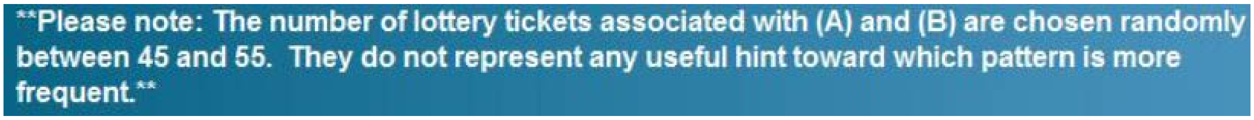
Sometimes the question had the form illustrated: The 3rd through 7th flip in the sequence are specified, the remainder are unspecified (designated "x"), and the participant must guess whether

[19] The wording of these questions—betting on which of the two patterns is "more frequent" out of the one million sets was a bad design choice, with a consequence that we did not intend: an agent whose beliefs are biased only by LSN should bet on the sequence believed to be more frequent, *regardless* of the payoff amounts. That is because, due to the law of large numbers, there is negligible uncertainty about which pattern occurred more frequently. The fact that we *do* observe sensitivity to the payoff amounts could be evidence of NBLLN, but it could also be due to choice noise whose magnitude is sensitive to the payoff levels, of the sort frequently modeled and found in empirical studies of random-utility maximization.

the sequence is more common with the 9th flip as a head or as a tail. A second variant, presented with equal probability, was in the mirror-image form, such as a choice between betting on (A) xHxTTTHHxx or (B) xTxTTTHHxx.

The first form of the question—which we call a “target-later” pair—allows us to test GF in the traditional “forward-looking” direction: Heads occurring *earlier* in the sequence make heads seem less likely to occur *later* in the sequence. Although virtually all of the existing evidence regarding GF is “forward-looking,” the logic of LSN—that small samples should have an unrealistically large chance of having 50% heads—also implies a “backward-looking GF”: Heads occurring *later* in the sequence should make heads seem less likely to have occurred *earlier* in the sequence. The second form of the question, a “target-earlier” pair, allows us to test for backward-looking GF.

The two options always offered a different number of lottery tickets. Because (A) and (B) are always equally likely, an unbiased agent would always strictly prefer the option that paid off more. In contrast to the vast majority of previous evidence, where GF is identified by participants’ choice of heads or tails with equal (real or hypothetical) payoffs, and hence no choice is actually erroneous, our setup can reveal unambiguous evidence of a bias if participants exhibit a systematic tendency in when they choose the low-payoff option. There were 6 different payoff possibilities for Option A / Option B: 55/45, 53/47, 51/49, 49/51, 47/53, and 45/55. In the instructions for this section, we told participants:

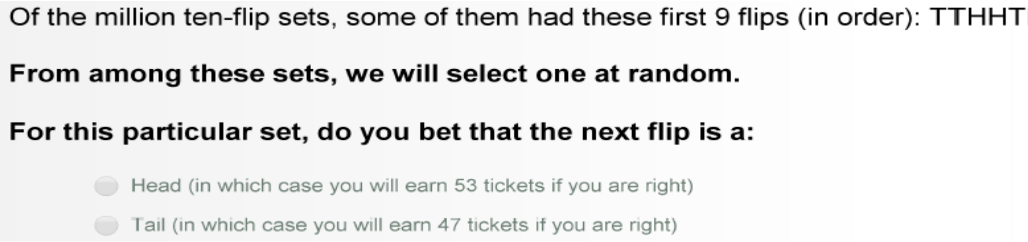


Please note: The number of lottery tickets associated with (A) and (B) are chosen randomly between 45 and 55. They do not represent any useful hint toward which pattern is more frequent.

Given 32 possible outcomes of the five specified coin flips, the two placements of the target flips, and the six possible payoff variations, there were 384 possible bets. Each participant saw 10 of these, presented randomly and independently.

Table 3 displays the fraction of times participants bet on heads as a function of the number of heads in the five known flips (expressed as the difference between the number of heads and the number of tails). In the full sample (column 1), the data do *not* seem consistent with a simple negative relationship, which would have been the most straightforward manifestation of GF. There is some, albeit noisy, evidence for that pattern in the target-later data (column 2), but not in the target-earlier data (column 3).

In Experiment 2, we made two changes in the betting questions relative to Experiment 1. First, we always showed the outcome of the first *n* flips and asked the participant to bet on the *n*+1st; each of the ten questions had a different value of *n* = 0,1,…,9. Unlike in the earlier experiment, we did not "hide" any of the flips (in order to try to make the task easier to understand and thereby reduce some noise), and we examined only forward-looking GF (in order to keep the survey length manageable, since we added a number of new questions in Experiment 2). Second, we changed from having the participant bet on which of the two patterns is "more frequent" to picking "one [of the sets] at random" and having the participant bet on the pattern for that set (in order to address the issue described in footnote 19).[20]



Of the million ten-flip sets, some of them had these first 9 flips (in order): TTHHTHTTT

From among these sets, we will select one at random.

For this particular set, do you bet that the next flip is a:

○ Head (in which case you will earn 53 tickets if you are right)

○ Tail (in which case you will earn 47 tickets if you are right)

Column 4 of Table 3 displays the fraction of times participants bet on the heads option as a function of the number of prior heads and the number of prior tails. As in Experiment 1, visual inspection of the table suggests that there is some evidence for GF, but the data are noisy. One source of noise is randomization of the payoffs. For example, due to lopsided randomization, respondents bet on a head in the first flip 58% (s.e. = 2.8%) of the time (not shown in the table), which is statistically distinguishable from 50%: 54.5% of respondents were offered higher payoffs for betting on Heads (see Appendix G.I.A.1.a.).

We turn to regression analysis to control for some of the sources of noise. Table 4 shows linear probability models where the dependent variable is a dummy for betting on the target flip

[20] When actually implementing this randomization to determine participants' payoffs, since the survey software we used had trouble with the full set of 512 possible 9-flip sequences, it drew instead from a randomly selected subset of 300. The reader may note that the survey question we asked is not really grammatical (it omits a question mark), but we show the question as it appeared.

being a head, and the key independent variable is the difference between the number of heads and the number of tails.[21] The control variables are indicators for the payoffs in favor of heads and, in Experiment 1, indicators for a target-later sequence and for the betting-on-heads option being listed on top. (In Experiment 1, the betting-on-heads option was randomly on top vs. bottom, but it was always listed on top in Experiment 2.) As expected, in both Experiment 1 (columns 1-3) and Experiment 2 (column 5), participants are more likely to bet on heads the higher the payoff for the betting-on-heads option.

Consistent with GF, the coefficient on the difference between heads and tails is negative in both experiments, but only statistically significantly so in the Experiment 2 data. The point estimates indicate that for every increase in the number of tails relative to heads, on average participants were 1.0 percentage point (s.e. = 0.8) more likely to bet on heads in Experiment 1 and 3.0 percentage points (s.e. = 0.5) in Experiment 2.

Although the difference in the coefficient across the target-later (column 2) and target-earlier data (column 3) is not statistically significant, it is more negative in the target-later data, which we interpret as consistent with the suggestive conclusion from Table 3 that forward-looking GF is stronger than backward-looking GF. This asymmetry contradicts the formal models of Rabin (2002) and Rabin and Vayanos (2010). If this hypothesized difference turns out to be real, one interpretation is that, in addition to believing in LSN, participants *also* exhibit a "causal-asymmetry" bias: Because (according to the logic of LSN) the earlier flip has a causal effect on what the later flip will tend to be, the earlier flip is mistakenly viewed as more predictive of the later flip than vice-versa. There is evidence for such an asymmetry in other contexts, e.g., people draw stronger inferences about a child's eye color from information about the mother's eye color than vice-versa (Tversky and Kahneman, 1980).

Next, we use our structural estimates from Section 3.2 to assess to what extent the apparent weakness of the evidence for GF in the betting data relative to the streak data might just reflect noisier behavior, as opposed to a weaker GF. To do so, we simulated how a Rabin-Vayanos agent

[21] In Appendix G.I.A.1.b, we report results from an alternative specification in which we replace the number of heads minus the number of tails with five indicator variables, one for each of the five flips closest to the target flip being a head. In the Experiment 1 data, the point estimates are negative for 4 of the 5 flip dummies (the exception being the 5th-closest flip), but only the coefficient on the 2nd-closest flip is statistically distinguishable from zero.

with the parameter values estimated from the streak data would behave in the betting task, taking into account the observed noisiness of the betting behavior. In each simulated dataset, we ran a linear regression of betting on heads on the difference between the number of heads and tails, including indicators for each of the payoff situations in the Experiment 2 simulations.[22] Columns 4 and 6 of Table 4 report the mean coefficients, averaged across the 1,000 simulations, for Experiments 1 and 2, respectively. In both experiments, the mean coefficient from the simulated data is larger in magnitude than the corresponding coefficient from the actual data—in Experiment 1, -0.018 versus -0.010, and in Experiment 2, -0.036 versus -0.030—but the point estimates from the simulated data lie comfortably within the 95% confidence intervals from the actual data. These results are suggestive that the GF is somewhat stronger in the streak data, but we cannot reject the hypothesis that the magnitude is the same in the betting data.

3.5. Summary: LSN and GF

Our findings point to a great deal of noise and heterogeneity in beliefs, with many participants not exhibiting LSN or GF and many participants confused by our elicitation techniques in ways that seem orthogonal to biases that seem economically important. Nonetheless, our data overall is supportive of LSN and GF, albeit weaker than in prior studies. Our evidence shows that people expect balance in small samples randomly drawn from a large sequence. While consistent with the psychology of LSN, our evidence is inconsistent with existing formal models; while LSN implies symmetry between forward-looking GF and backward-looking GF, we find no evidence for backward-looking GF.[23]

[22] For each of Experiment 1 and 2, we used the parameter values estimated from the respective streak data. To estimate the noisiness of the betting behavior, for each of the six possible payoff situations in each experiment, we calculated the "mismatch rate" at which participants' bets differed from what the Rabin-Vayanos agent would do. Across the six payoff situations, the mismatch rates in Experiment 1 ranged from 39% to 47%, and in Experiment 2, from 26% to 40%. Then we simulated 1,000 datasets by taking what the Rabin-Vayanos agent would have done when faced with the participants' options, and randomly switching the choice according to the respective mismatch rates for the given payoff scheme. In the regressions, since neither of the other variables we manipulated—target earlier vs. target later and whether the heads option is listed on top—matters in the Rabin-Vayanos model, we do not include them as controls. We similarly omitted controls for the payoffs in Experiment 1 because, given our wording there (see footnote 19), these also should not matter for the Rabin-Vayanos agent. The standard errors are calculated as the square-root of the sum of two terms: The mean of the squared standard errors across the 1,000 simulations, and the variance of the estimated coefficient across the 1,000 simulations; the second term takes into account the uncertainty from the simulation.

[23] Below we will show another inconsistency: we find GF in cases where no existing formal models predict it. Also, as we expected in this context of coin flips, we did not see much prevalence of the seeming opposite of GF, the "hot-

4. Bin Effects

Next, we examine the evidence in our data for "bin effects," a term we use to refer to how participants' beliefs are influenced by the bins into which they are asked to categorize different outcomes. In Section 4.2, we develop a simple theoretical framework to relate the bin effects in our experiment to support theory, and we derive some implications that we use in Section 5 to interpret participants' histogram beliefs.

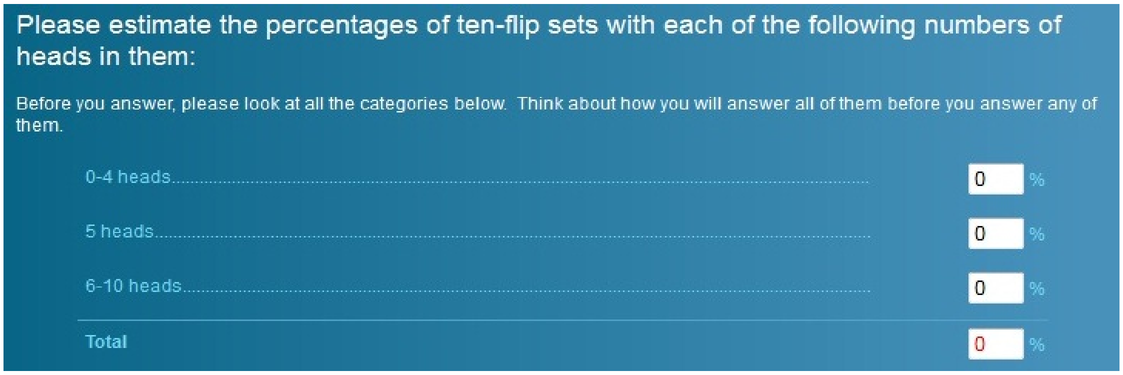
4.1. Evidence on Bin Effects

To elicit participants' beliefs of full subjective probability distributions over different possible outcomes, we asked participants to type in a number between 0 and 100 for each member of a set of exclusive and exhaustive ranges of outcomes. The screen showed the sum of the percentages, with a sum of 100% required before they could continue to the next screen. Following Haran, Moore, and Morewedge (2010), we refer to this type of question as Subjective Probability Interval EStimates (SPIES). As part of the general instructions for each experiment, we gave participants training for the SPIES interface by asking them to estimate the percentage of the population in the United States composed of each of six major racial groups (White, Hispanic, Black, Asian, Native American, and multiracial).

For a sample of size ten, in both experiments we elicited participants' histogram beliefs with four types of (randomly ordered) questions that binned the possible outcomes in different ways. The first three types were SPIES. One of these asked the participant to estimate, among the million samples of size ten, the frequency of 0-4, 5, and 6-10 heads[24]:

hand fallacy" whereby people expect streaks to continue. A long tradition in psychology posits GF as the foundation for the hot-hand fallacy: because people expect fewer streaks than randomness actually generates, in situations where some narrative about streaks makes sense to people, they will come to believe in more streaks than exist. See Rabin and Vayanos (2010) for a formalization of this argument. Finally, note that our methods do not run afoul of criticisms by Miller and Sanjurjo (2016) of other GF and hot-hand-fallacy evidence.

[24] Unfortunately, for this one question a bug in the survey format resulted in half of participants being asked to respond with the percentage of tails even though the question referred to the numbers of heads. Given that the two interpretations are meaningfully identical, we suspect that our mistake had little effect on participants' answers.



Please estimate the percentages of ten-flip sets with each of the following numbers of heads in them:

Before you answer, please look at all the categories below. Think about how you will answer all of them before you answer any of them.

0-4 heads .... 0 %

5 heads .... 0 %

6-10 heads .... 0 %

Total 0 %

The other two were SPIES with the five categories 0-3, 4, 5, 6, and 7-10 heads, and one with the eleven categories 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 heads. The fourth type of question was a set of 11 questions (in random order), separating out each of the possible outcomes for a sample of size 10. For example: "What percentage of ten-flip sets include exactly 4 HEADS and 6 TAILS?" Each of these questions effectively creates a 2-bin elicitation, the focal outcome and everything else (in the example, "4 heads" and "not 4 heads").[25]

For each participant, we randomized (by question block) whether the question was framed in terms of heads or tails; whether the bins were labeled in terms of the number of "heads" out of *N* flips or the number of "tails"; and whether the screen asked participants to guess the frequency in increasing or decreasing order of heads. As we discuss in Appendix G.I.D.g-i, a few of these counterbalancing variations are statistically significant, but none substantively influence the results we report below.[26] Therefore, our analyses pool the data across the randomizations.

For Experiment 1, the mean histogram beliefs for each of the four types of questions are

[25] As we also note below, this type of question may not have produced the same sort of bin effects as elsewhere because it may have focused participants' attention more on the particular outcome we asked about (e.g., 4 heads) than if we had asked the participant to fill in a histogram that listed both that outcome and the complementary outcome.

[26] In particular, in Experiment 1, there evidence that in some histograms, participants put higher weight on the middle bin if the bins were listed in increasing order of heads; and in Experiment 2, there is evidence in some histograms of the opposite effect (Appendix G.I.D.i). As a more consistent pattern, in Experiment 1 (but not Experiment 2), we find that on average across all the histograms, participants put higher weight on the first bin that was listed than on the last bin that was listed (Appendix G.I.D.g). This pattern suggests that in Experiment 1, participants may have begun filling in the SPIES from top to bottom with relatively large weights, and then started responding with smaller weights toward the bottom when they got close to hitting the constraint of 100% total probability. Recall that we randomized whether, in any particular histogram elicitation, the histogram was framed to a participant in one of four ways: (1) 0-3, 4, 5, 6, and 7-10 heads, *or* (2) 0-3, 4, 5, 6, and 7-10 tails, *or* (3) 7-10, 6, 5, 4, and 0-3 heads, *or* (4) 7-10, 6, 5, 4, and 0-3 tails. In order to have our data accurately reflect participants' tendency to put higher weight on the earlier bins listed, we pool responses from the first category of each of these four and call it "0-3 heads," the second category of each and call it "4 heads", and so on.

shown in Figures 5, 6, 7, and 8; and for Experiment 2 in Figures 9, 10, 11, and 12. Note that Figures 8 and 12 show the beliefs when each outcome was elicited separately; in these histograms, the beliefs sum up to 345% in Experiment 1 and 165% in Experiment 2.[27]

We find strong bin effects for the sample size of ten flips, in the direction that more bins leads to more compression of the weights assigned to the bins. For example, for the outcome 5 heads out of 10 flips (whose true probability is 25%), mean beliefs are 39%, 36%, 28%, and 20% in going from 2-bin to 3-bin to 5-bin to 11-bin elicitation. (The pairwise p-value from the 2-bin to the 3-bin elicitation is 0.34, but below 0.0001 for the other two.) In Experiment 2, the effects are smaller, but the results are consistent with bin effects, yielding mean beliefs of 34%, 33%, 32%, and 28%. (The pairwise p's for the first two are 0.22, 0.29, but less than 0.0001 for the last pair.[28]) At the individual level, 40% of participants in Experiment 1 assigned weakly lower probability to the middle bin as the number of bins increases from 2- to 11-bin elicitations, compared with only 10% who have the opposite pattern of weakly higher probability; the corresponding fractions of participants in Experiment 2 are 33% and 20% (Appendix G.I.A.2.b.v).

For a sample of size one thousand, we have two tests for bin effects, using data from four SPIES elicitations that participants faced in a random order. Our first test uses two of these questions. One question mirrored Kahneman and Tversky's (1972) eleven-bin elicitation of the histogram for a sample of size one thousand, with the events binned according to the following number of heads: 0-50, 51-150, 151-250, 251-350, 351-450, 451-549, 550-649, 650-749, 750-849, 850-949, and 950-1000. A second question had 3 categories: 0-450, 451-549, 550-1000. The results from Experiment 1 are shown in Figures 13 and 14, and for Experiment 2 in Figures 15 and 16.

In both experiments we find strong evidence of bin effects, again consistent with more compression when the number of bins is larger. Going from the 3-bin to the 11-bin histogram, the

[27] Indeed, we expect that due to bin effects, the sum of participants' probabilities will exceed 100% when each of at least three probabilities was elicited separately. Let $P(E|\mathcal{E})$ be the reported probability for event E given partition of state space $\mathcal{E}$. Then we should expect $P(x|\{x\},\{y,z\}) + P(y|\{y\},\{x,z\}) + P(z|\{z\},\{x,y\}) > P(x|\{x\},\{y\},\{z\}) + P(y|\{x\},\{y\},\{z\}) + P(z|\{x\},\{y\},\{z\})$. We are aware of one previous test of whether such probabilities sum to greater than 100% (Teigen, 1974), which found support for that implication in unincentivized belief elicitations regarding samples of size 5 and 10.

[28] The corresponding evidence from median beliefs is similar: In Experiment 1, the four respective median beliefs are 36%, 39%, 25%, and 14% (where the first two are not statistically distinguishable, but all other differences are highly significant), and in Experiment 2, they are 35%, 32%, 30%, and 25% (pairwise p's: 0.06, 0.47, and <0.0001).

mean probabilities assigned to the middle bin (451-549 heads, which has true probability 99.8%) were 40% and 19% in Experiment 1 and 41% and 34% in Experiment 2.[29] At the individual level, 95% of participants in Experiment 1 assigned weakly lower probability to the middle bin in the 11-bin histogram than in the 3-bin histogram, compared with only 18% who assigned weakly higher probability; in Experiment 2, the numbers were 69% and 49% of participants.

Our second test draws on two further questions. One of them was an 11-bin histogram, with the bold highlighting the difference from the earlier question: 0-50, 51-150, 151-250, 251-350, 351-**480, 481-519, 520**-649, 650-749, 750-849, 850-949, and 950-1000. The other is a 3-bin histogram with categories 0-480, 481-519, and 520-1000. In Experiment 1, we cannot compare these two because we only asked the 3-bin question (Figure 17). For Experiment 2, the results are displayed in Figures 18 and 19. We again find evidence for bin effects, with more compression when there are more bins: The mean probabilities assigned to the middle bin (481-519 heads, which has true probability 78%) were 30% and 36%.[30] 66% of participants assigned weakly higher probabilities in the 3-bin histogram than the 11-bin histogram, compared with 46% who assigned weakly lower probabilities.

4.2. A Simple Theoretical Framework for Bin Effects

We now develop a simple theoretical framework for bin effects allowing us to (a) relate the bin effects observed in our experiments to support theory (Tversky and Koehler, 1994), and (b) derive implications that we will use in Section 5 to make inferences about participants' distributional beliefs "disentangled from" bin effects.

To establish our notation, each possible outcome of a set of coin flips, such as the outcome "8 heads out of 10," is called a state of the world, denoted $\omega \in \Omega$, where Ω is the set of all possible states. A subset of Ω, such as "8, 9, or 10 heads out of 10," is called an event and is denoted $E \subseteq \Omega$. Each of the histograms we presented to participants represents a particular "binning" of the state space. Formally, a binning $\mathcal{E}$ is a set of mutually exclusive events that jointly cover the state space. The agent's "stated belief" about the probability of any event $E \in \mathcal{E}$ may depend on the binning $\mathcal{E}$.

[29] The corresponding median probabilities are 40% and 11% in Experiment 1 and 40% and 28% in Experiment 2.

[30] The corresponding median probabilities are 25% and 34%.

Support theory says that there exists a function $s(\cdot)$ *defined independent of the binning* that maps any possible event into a strictly positive number and satisfies two assumptions:

A1. $P(E|\mathcal{E}) = \frac{s(E)}{\sum_{F \in \mathcal{E}} s(F)}$,

A2. For any mutually exclusive events E' and E'', $s(E'\,) + s(E'') \geq s(E' \cup E'')$.

The function $s(\cdot)$ is called the "support function," and A1 formalizes the sense in which it represents the strength of belief in an event. A2 states that the support function can be subadditive. If A2 always holds with equality, then the support function $s(\cdot)$ would be additive; it would simply represent a standard subjective probability (and it would *equal* a subjective probability if it were rescaled so that $\sum_{F \in \mathcal{E}} s(F) = 1$). With a strict inequality, A2 can accommodate the evidence from Tversky and Koehler (and similar studies) that asking people the likelihood of getting various cancers and adding up those likelihoods leads to a larger number than asking people the likelihood of getting cancer.[31]

We extend this framework to allow us to study biases relative to correct beliefs. We assume that the agent has "root beliefs" that satisfy the standard rules of probability. These root beliefs may be biased (e.g., by NBLLN), and we would like to infer what the root beliefs are so that we can compare them with the correct beliefs, but we observe stated beliefs that are distorted by bin effects. We denote the root belief about event E by $r(E)$. We denote the stated belief, which depends on the binning $\mathcal{E}$, by $P(E|\mathcal{E})$ as above.

Our framework for studying bin effects is defined by two assumptions, which are closely related to A1 and A2. The first is:

$\mathbf{A1'}$. $P(E|\mathcal{E}) = \frac{g(r(E))}{\sum_{F \in \mathcal{E}} g(r(F))}$, where $g(\cdot)$ is a strictly increasing, positive-valued function.

Assumption A1$'$ is equivalent to A1 together with the assumption that the support of an event is a fixed, increasing function of the agent's root belief: $s(E) = g\big(r(E)\big)$ for all $E \subseteq \Omega$. Some assumption along the lines of A1$'$ is necessary if bin effects are operating on underlying (stable)

[31] Ahn and Ergin (2011) provide axiomatic foundations for support theory.

root beliefs. Assumption A1$'$, however, has a strong and immediate implication: The stated-belief odds of any two events are independent of the binning.

Proposition 1: Assume A1$'$, and fix any two events and two binnings that include those events, $E, E' \in \mathcal{E}, \mathcal{E}'$. Then $\frac{P(E|\mathcal{E})}{P(E'|\mathcal{E})} = \frac{P(E|\mathcal{E}')}{P(E'|\mathcal{E}')}$.

(All proofs are in Appendix E.) In each of our experiments, there is only one available direct test of this prediction: We can use participants' reported beliefs to calculate the implied odds of 5 heads versus 4 heads out of 10 in the 5-bin histogram, and we can compare these to the implied odds of 5 heads versus 4 heads in the 11-bin histogram.[32] (In conducting this test, we analyze participants' mean beliefs, and instead of using the reported belief of 4 heads, we use the average of the reported beliefs of 4 heads and 6 heads.) In Experiment 1, these odds are 1.41 and 1.60 in the 5- and 11-bin histograms, respectively; in Experiment 2, they are 1.76 and 1.89. In both experiments, the difference between the odds approaches statistical significance (p = 0.12 in Experiment 1 and p = 0.054 in Experiment 2, from a permutation test). Nonetheless, we view Assumption A1$'$ as a reasonable enough first approximation. Indeed, we are informally relying on it when (in Section 5) we extrapolate from the amount of compression we observe in one histogram to the amount of compression that could plausibly be attributed to bin effects in another histogram. However, our formal results below rely on a weaker assumption that allows the $g(\cdot)$ function to depend on the binning:

$$\mathbf{A1''}. P(E|\mathcal{E}) = \frac{g_{\mathcal{E}}(r(E))}{\sum_{F \in \mathcal{E}} g_{\mathcal{E}}(r(F))}, \text{ where } g_{\mathcal{E}}(\cdot) \text{ is a strictly increasing, positive-valued function.}$$

Unlike Assumption A1$'$, Assumption A1$''$ does not imply that the odds must be equal. Assumption A1$''$ immediately implies the following proposition:

Proposition 2: Assume A1$''$, and fix any two events, $E, E' \in \mathcal{E}$. Then: $\frac{P(E|\mathcal{E})}{P(E'|\mathcal{E})} > 1$ if and only if

[32] When we designed the experiment, we had not yet formulated this framework for bin effects. As a result, we did not build in tests of this assumption or other implications of the framework.

$\frac{r(E)}{r(E')} > 1$; and $\frac{P(E|\mathcal{E})}{P(E'|\mathcal{E})} = 1$ if and only if $\frac{r(E)}{r(E')} = 1$.

The proposition says that, for a fixed binning, the reported belief for event E is greater than the reported belief for event E' if and only if the root belief for E is greater than the root belief for E'. This result has two useful implications. First, even though bin effects distort the magnitudes of reported beliefs, we can use the reported beliefs to infer the *ordering* of likelihoods assigned to events by the root beliefs, i.e., which event is judged to be most likely, second-most likely, etc. Second, there is a special situation in which we can infer the magnitude of root beliefs: When the reported beliefs are equal to each other. Specifically, if there is some binning under which a person reports equal beliefs for every bin, then we can infer that the person's root beliefs also assign equal weight to each bin.

We obtain an additional useful implication when we add Assumption A2$'$:

A2$'$. For any $\mathcal{E}$, $g_{\mathcal{E}}(\cdot)$ is a concave function.

Assumption A2$'$ is essentially equivalent to A2, given Assumption A1$''$. The evidence of greater compression when there are more bins from Section 4.1 is consistent with Assumption A2$'$.

Proposition 3: Assume A1$''$ and A2$'$, and fix any two events, $E, E' \in \mathcal{E}$. If $\frac{P(E|\mathcal{E})}{P(E'|\mathcal{E})} > 1$, then $\frac{r(E)}{r(E')} > \frac{P(E|\mathcal{E})}{P(E'|\mathcal{E})}$.

The proposition says that, for a fixed binning, for any two events for which the reported beliefs differ, the root-belief odds of E to E' are more extreme than the reported-belief odds of E to E'. Intuitively, since the reported beliefs are a compressed version of the root beliefs, we know that the root beliefs are more extreme than the reported beliefs. This result implies that there is another special situation in which we can draw stronger inferences about the root beliefs: When the true probabilities of each bin are equal to each other. Because observing a person stating one event is more likely than another in a given binning tells us the direction of the root beliefs, we know the root beliefs are distorted in that direction relative to the true (equal) probabilities. We will rely on the implications of Propositions 2 and 3 in the analyses in the next section.

5. Biased Beliefs about Distributions of Outcomes

In this section, we turn to evaluating people's histogram beliefs. In particular, we examine whether participants exaggerate the likelihood of extreme proportions as predicted by NBLLN and whether they exaggerate the likelihood of the mean proportion as predicted by exact representativeness.

5.1. Replicating Kahneman and Tversky's (1972) Evidence for Sample-Size Neglect

We begin by comparing our evidence to Kahneman and Tversky's (1972) evidence for sample sizes of ten and a thousand, as shown in Table 5. Our results differ from theirs in some of the details. For example, for the sample size of ten in Experiment 2, participants' median beliefs not only overweight the extremes of the distribution but also overweight the middle bin of 5 heads. As another example, our participants in Experiment 1 put more probability mass in the bins presented to them earlier (as per footnote 26), e.g., for the sample size of 1000, they put higher median probability (5.0%) on 0-5% heads than on 95-100% heads (2.5%). Except for these extreme bins, however, we cannot reject the hypothesis that the mean probability for a given bin is independent of whether the coin was flipped ten times or one thousand times. Indeed, Figures 20 and 21 show the same "sample-size neglect" phenomenon for median beliefs in our experiments as Kahneman and Tversky found, and Figures 22 and 23 show the same pattern for mean beliefs. Overall, our findings qualitatively replicate Kahneman and Tversky's.[33]

Since the number of bins is held constant across the sample sizes—of ten, one hundred, and one thousand in Kahneman and Tversky's evidence, and ten, one thousand, and one million in ours—this evidence suggests that participants' "root beliefs" over the proportion of heads are essentially invariant to sample sizes in the range of 10 to 1 million. This fact in itself constitutes evidence of NBLLN, since the law of large numbers should imply that the distribution becomes far more concentrated around 50% as the sample size increases.

[33] As described in Appendix G.I.D.b, we also find that individuals whose distributions are more spread out for samples of size 10 also tend to be spread out for samples of size 1,000: The individual-level correlation between the standard deviations is 0.67 for Experiment 1 and 0.42 for Experiment 2.

Taken at face value, these data can be interpreted—as, for instance, Benjamin, Rabin, and Raymond (2016) do—as evidence that participants overweight the extremes of the distribution for all of the sample sizes, even the sample size of 10. This conclusion, however, is not necessarily warranted because it confounds the overweighting of extreme outcomes *per se* implied by NBLLN with bin effects that bias all bins toward equal weight, as we discuss next.

5.2. NBLLN

We now turn to disentangling the evidence for NBLLN, as manifested by overweighting the extremes of the distribution, from bin effects. The key worry is that bin effects lead in general to overweighting of low-probability events, and most elicitations of histogram beliefs ask in a way where extreme outcomes are the lowest-probability outcomes. Instead, we focus on our histograms binned such that the extreme outcomes had at least as high probability as less extreme outcomes. Our framework from Section 4.2 then allows us to draw inferences about the root beliefs.[34]

Consider first the sample size of ten. Recall that if we examine a binning such that people's stated belief about the likelihood of each bin is equal, then we can infer that the root beliefs are also equal. While none of our histograms achieve this exactly, the 5-bin histogram of 0-3, 4, 5, 6, and 7-10 heads, with true probabilities of 17%, 21%, 25%, 21%, and 17%, comes closest. Figures 6 and 10 reveal that participants' mean beliefs with this binning are approximately correct—18%, 22%, 28%, 18%, and 14% in Experiment 1 and 16%, 18%, 32%, 18%, and 16% in Experiment 2.[35] We interpret this evidence as indicating that, once bin effects are accounted for, there is little evidence of overweighting the extremes in samples of size 10. Indeed, as we return to below, there is some overweighting of the middle bin.

For samples of size a thousand, by contrast, the evidence for overweighting the extremes is strong. Although (as shown in Figure 24) we generated *under*weighting of the extremes with the binning 0-499, 500, and 501-1000 heads, where mean beliefs on 500 heads were 19% rather than the correct 2.5%, the overweighting of the middle is consistent with bin effects, and indeed this question was designed to show the power of the bin effects. A first indication of the overweighting

[34] For brevity, we focus primarily on discussing mean beliefs, but Appendix G.I.A.2.a shows that results for median beliefs are nearly identical.

[35] Even though our argument here treats the mean beliefs as essentially equal to the true probabilities, we note that we can statistically reject that null hypothesis for many of the mean beliefs. In Experiment 1, the respective *p*-values are 0.45, 0.40, 0.03, 0.06, and 0.0006; and in Experiment 2, 0.05, < 0.0001, < 0.0001, < 0.0001, and 0.01.

of extremes comes from the symmetric 3-bin histograms. Figures 17 and 19 show that when the middle bin is 481-519, the average assigned to each of the two non-middle bins is 32% rather than the true probability of 11%. Although bin effects of a plausible degree could possibly explain this seeming overweighting of tails, the 3-bin treatment with a middle bin of 451-549, shown in Figures 14 and 16, invites no such interpretation. The average beliefs for the two extremes are 30%, rather than the true 0.01%. It is implausible that participants would have had the right beliefs in the absence of bin effects; transforming such probabilities 3,000-fold goes against the evidence from both our experiment and prior research.[36] Moreover, in a 3-bin histogram with essentially equal true probabilities, 0-493, 494-506, and 507-1000 heads (Figure 25), participants' mean beliefs overweight the extremes: Although the true probability of 494-506 heads is 32%, participants' mean probability is 29% (lower than 32% at $p = 0.05$), and their median probability is 25%. From these results, we see a small but clear bias toward overweighting the extremes.

Interesting additional evidence comes from the only elicitation we did with asymmetric binning: In Experiment 2, we asked participants the proportion of time the samples would come out 0-909, 910, and 911-1000 heads. Although this was not the original plan of focus for this data, we realized that this question has a useful and interesting property: Getting between 0 and 909 heads is of course nearly certain, but the true probability of the far extreme bin (911-1000 heads), which is 1.11×10^{-172}, is *lower* than the true probability of the intermediate extreme bin (910 heads), which is 1.01×10^{-171}. Yet Figure 26 shows participants' mean belief on the far extreme, 13.5%, is higher than on the intermediate extreme, 7.5%.[37] This reversal cannot be explained by bin effects and is consistent with overweighting the tails.

[36] To calibrate what is plausible, consider again the results shown in Figures 8 and 12, where participants were asked to assess frequency of the bins 0, 1, 2, 8, 9, and 10 in isolation. The actual probabilities ranged from 0.1% to 4.4%, but participants reported beliefs between 18% and 28% in Experiment 1 and between 4% and 12% in Experiment 2. Such two-bin elicitations should lead to bigger bin effects on low probability categories than for the example in the text, but led to maximal beliefs of 28% and 12% in the two experiments, and maximal distortion of categories with probabilities 0.1% were to 18% and 4%. It is thus unlikely that bin effects would distort the 0.1% to 30% in the example in the text. Of course, in principle we should not be comparing distortions of true probabilities, but rather the beliefs taking the judgmental biases into account. This too does not seem a problem here, given that we induce roughly accurate beliefs for sample sizes of 10 once bin effects are controlled for. Similarly, for the histogram 0-480, 481-519, and 520-1000 heads, it seems unlikely that participants' reports of near-equal probabilities of bins that actually have probabilities 11%, 78%, and 11% are due solely to bin effects; the isolated-category bins of 3 heads and 7 heads, each with true probability 11%, in fact induce reports about 30%, but in these cases we'd expect to see stronger bin effects than in the 3-bin example because it is with salient elicitation and two bins.

[37] Similarly, the median beliefs are 6.3% and 1.0%, respectively. Moreover, 63% of respondents reported beliefs smaller than 1% for 910 heads, which suggests that participants are willing to report such small numbers when their beliefs warrant them.

A final source of evidence for NBLLN net of binning effects comes from the sample size of a thousand and the 5-bin histogram with nearly equal true probabilities, 0-487, 488-496, 497-503, 504-512, and 513-1000 heads. Because the true probabilities are equal, the direction of deviation in beliefs away from equality should reveal the direction of bias in root beliefs net of bin effects. As shown in Figure 27, we find a "W pattern"—overestimation of the probability of the middle bin *and* of the extreme bins—consistent with a combination of overweighting both the middle bin and the extreme bins. This pattern for mean beliefs, however, results from roughly half of the participants (48%) overweighting both extreme bins and roughly half (50%) overweighting the middle bin, with very few (5%) overweighting both (Appendix G.I.B.1.b.ii.). We view this third source of evidence as suggesting that both overweighting the middle bin and overweighting the extreme bins are common biases, although one or the other may dominate for different individuals.

Turning to beliefs about sample sizes of one million, in the first of two conditions in Experiment 2, we elicited beliefs about a 3-bin histogram with equal true probabilities: 0-499784, 499785-500215, and 500216-1000000 heads. We were shocked to find that participants' mean and median beliefs in this case were essentially correct (Figure 28). However, at the individual level, only 14% of participants gave roughly equal weight of 30-36% to the three bins, compared with 44% who gave higher probability to the middle bin than both extreme bins and 44% who gave higher probability to both extreme bins (Appendix G.I.C.1.b). Thus, at the individual level, the correct answer is rare, while NBLLN and overweighting the middle bin are equally common. The second histogram for a sample size of 1 million had 5 bins, 0-499579, 499580-499873, 399874-500126, 500127-500420, and 500421-1000000, with the bins also chosen to have equal probabilities. As shown in Figure 29, in this case we see the W pattern seen earlier for 5-equally-likely-bin elicitations: Compared to the true probability for the middle bin of 19.9%, participants' mean belief is 25.6%; and compared to the true probability for each extreme bin of 20.1%, participants' mean beliefs for the left and right extreme bins are 24.4% and 24.6%. As with the sample size of 1,000, however, the means mask heterogeneity: 40% of the participants overweight both extreme bins, and somewhat more than half (56%) overweight the middle bin, with very few (5%) overweighting both (Appendix G.I.C.1.b.ii). Although overweighting the middle bin seems bigger than NBLLN in this case, the evidence clearly indicates the presence of both biases.

Although we were surprised that people did not put much more weight on the extremes than they did, the sample size of 1 million also supports NBLLN.

5.3. Exact Representativeness

Much of the evidence mentioned in the previous subsection not only points toward overweighting the extremes of the distribution but also overweighting the middle bin. Participants seemed to overweight the middle bin slightly in the 5-bin histogram for the sample size of ten, and seemed to overweight both the middle and the extremes in the 5-bin histogram for sample sizes of a thousand and a million. Figures 7 and 11 show evidence of overweighting the middle bin in the 3-bin histogram of 0-4, 5, and 6-10 heads: the true probability of the middle bin is 25%, but the mean belief for the middle bin is 36% in Experiment 1 and 33% in Experiment 2. While we tended to see proper weighting or underweighting of the middle bin in other 3-bin histograms on average, substantial fractions of the experimental participants overweighted it. In these cases, our theoretical framework thus implies that the root beliefs exhibit a bias toward overweighting the middle bin.

Can the overweighting of the middle bin be explained by GF, which causes people to expect too much alternating between heads and tails and therefore too much mass in the middle of the distribution? Or is there an additional bias—which, following Camerer (1987), we call "exact representativeness" (ER)? To get at this question, we use our structural estimates of GF from Section 3.2 to test whether the overweighting of the middle bin that we observe can be explained by the amount of GF in the streak data.[38] Specifically, for each histogram from Experiment 1, we simulate what a Rabin-Vayanos agent with our estimated Experiment-1 parameter values would believe; and similarly for the Experiment-2 histograms using the estimated Experiment-2 parameter values. Each histogram figure plots the resulting Rabin-Vayanos beliefs as triangles. Across all the histograms, the Rabin-Vayanos beliefs are quite close to the correct probabilities,

[38] In our experimental design, we included the 3-bin histogram, 0-480, 481-519, and 520-1000 heads, with the intention that comparing it to the other 3-bin histogram, 0-450, 451-549, and 550-1000 heads, might provide "smoking gun" evidence of ER in the form of a sort of "conjunction fallacy." If the weight assigned to 481-519 heads were *higher* than the weight assigned to 451-549 heads, then ER must be at play, inducing people to put higher weight on a subset if that subset more closely resembles the underlying 50% expected outcome. However, in both experiments, participants' mean probability of 451-549 heads was higher than their mean probability for 481-519 heads: in Experiment 1, the respective means were 39.6% versus 35.6% (p = 0.009); and in Experiment 2, 41.1% versus 36.3%. This rules out an extreme form of ER, but is consistent with the other evidence in favor of ER.

so that the conclusion above that (for all three sample sizes) participants' root beliefs overweight the middle bin relative to the true probabilities is also evidence of overweighting relative to the Rabin-Vayanos beliefs.[39] Overall, participants' histogram beliefs overweight the middle bin to a greater extent than can be explained by the amount of GF they exhibit in their streak beliefs.

6. Can Sequence and Histogram Beliefs Be Jointly Rationalized?

The bin effects that we observe seem clearly at odds with any internally consistent model of coin flips that experimental participants might have. In this Section, we put aside the bin effects, and we ask whether participants' sequence beliefs could be internally consistent with their root beliefs about the proportions of heads (i.e., their histogram beliefs net of bin effects). Our integrated design—with participants responding to different questions about the very same set of flips—allows us to test internal consistency of beliefs in a much stronger way than is usually possible. Specifically, as long as participants trusted our instructions, they could not have thought that the coin flips were generated by a different process when we asked about streaks as when we asked about histograms.

We conduct two tests. First, we compare participants' sequence beliefs with their histogram beliefs for the relative likelihood of two events that we asked about in both elicitation formats: 9 heads out of 10 relative to 10 heads out of 10. In the sequence questions (Figures 2 and 3), when asked how frequently 9 heads is followed by a head, participants in both experiments reported beliefs (per GF) that HHHHHHHHHH is roughly half as likely as HHHHHHHHHT. And (given GF) it is surely the case that participants think that the nine other ways to get 9-out-of-10 heads are at least as likely as HHHHHHHHHT. Therefore, if participants' histogram beliefs corresponded to their sequence beliefs, they would think that 9 out of 10 heads should be at least 20 times more likely than 10 out of 10 heads. Yet when asked separately about each possible number of heads out of 10 flips (Figures 5 and 9), they reported in Experiment 1 that 9 heads is only 2.3 times more likely than 10 heads—and in Experiment 2, only 1.6 times. These responses

[39] The one exception is the 5-bin histogram for the sample size of 10 (Figures 6 and 10), where participants' mean belief about 5 heads of 28.1% in Experiment 1 was larger than the true probability of 25% (p = 0.03) but not much different from the Rabin-Vayanos belief of 27.8% (p = 0.86). In Experiment 2, the mean belief of 32.1% on the middle bin is larger.

are influenced by the binning, of course, so given our evidence that participants' root beliefs are roughly calibrated correctly for a sample size of 10 (net of bin effects), we might instead consider the true ratio of the probabilities, which is 10. Even this true ratio, however, is far smaller than the ratio (of at least 20) implied by the sequence beliefs, indicating that participants' beliefs are not internally consistent.[40]

The second test addresses "parameter uncertainty" about the bias of the coin, which is the most plausible internally consistent theory that could reconcile a belief in GF (i.e., negative autocorrelation) in sequences with NBLLN (i.e., putting too much probability on extreme sample proportions) in histograms. Specifically, people might think that any given set of flips is generated by a coin whose probability of heads is randomly drawn—equal to 50% on average but not on any particular occasion—and then the flips themselves are negatively autocorrelated. Such an agent would exhibit GF in the sequence data (since any sequence is generated by a coin with a fixed probability of heads), but the agent's histogram beliefs would put too much weight in the extremes of the distribution due to a belief that some of the coins are heavily biased toward heads or tails.

We conducted Experiment 2 partly in order to implement our second test, which we formulated after Experiment 1. The test uses the nine questions—three each for the sample sizes of ten, a thousand, and a million—discussed above in Section 3.3, regarding the probability that a randomly chosen flip in a sample is a head, conditional on 1, 2, or 5 other randomly chosen flips being heads. The basic idea is that if an agent's histogram beliefs are overly dispersed, then if one flip is *randomly chosen from a fixed sample* and turns out to be a head, then the agent should believe that the sample is more likely to be a majority-heads sample than if the histogram beliefs were correct. Consequently, if another flip is randomly chosen from the same fixed sample, an agent with an internally consistent model should believe that that flip's likelihood of being a head is greater than 50%. And if additional flips are drawn and come up heads, then the agent should believe that the sample is more and more tilted toward heads and should believe that the next flip's likelihood of being a head is even higher. This logic goes through regardless of what the agent

[40] Another test along similar lines is to compare the probability of 10 out of 10 heads implied by the streak data with the reported probability in the histogram data. According to the mean beliefs in the streak data, Pr(H) × Pr(H|H) × Pr(H|HH) × … × Pr(H|HHHHHHHHH) = 0.005% in Experiment 1 and 0.01% in Experiment 2. This is *much* smaller than the reported probability of 10 heads out of 10 in Figures 5 and 9 of 2.8% and 2.4%, respectively, as well as much smaller than the true probability of 0.1%.

believes about the random process of coin flips that generated the fixed sample—that is, regardless of what the agent may believe about the correlation structure across flips.

To formalize this idea, let N be the size of the fixed sample of flips (10, 1000, or 1 million in our experiment), and let $\pi(M, N)$ denote the probability that an additional randomly chosen flip from the sample will be a head when $M < N$ flips have been randomly chosen from the sample and all turn out to be heads.

Proposition 4: Suppose the bias of the coin is not known to be 50% but is instead drawn from ν, a (continuous or discrete) nondegenerate distribution on [0,1] that has mean μ_ν. Then $\pi(1, N) > \mu_\nu$, and $\pi(M, N)$ does not depend on N and is strictly increasing in M.

Contrary to these predictions, as shown in Figure 4 and also discussed above in Section 3.3, for all three sample sizes, participants' mean beliefs are *smaller* than 50% and *decreasing* in the prior number of randomly drawn flips that were heads. While participants' beliefs are consistent with a psychologically plausible form of LSN, they cannot be rationalized with participants' histogram beliefs by a model of parameter uncertainty.

7. Robustness and Additional Analyses

In accordance with our pre-specified analysis plans, we conducted additional analyses to shed further light on our results and to probe their robustness.

Order effects: Inferences and fatigue. When we designed our experiments, we took steps—such as explicitly telling participants that some of the numbers in the questions were chosen randomly—to deter participants from drawing inferences from the questions about what answers they should give. To test if participants might have been drawing such inferences despite our efforts, we examined, for some of our belief elicitations, whether mean beliefs varied depending on *when* in the experiment the elicitation was conducted. The idea is that, if participants' responses were influenced by inferences from earlier questions, then the same question asked later in the experiment might elicit a different answer. Specifically, we examined the streak questions for all three sample sizes in Experiment 2.[41] We found no evidence of systematic differences in responses depending on whether these questions were asked in the first, middle, or last third of the experiment (Appendix Tables H.4.A-C, column 5).

If different participants drew different inferences over the course of the experiment, or if participants' responses were influenced by fatigue later in the experiment, then participants' responses to a given question might be noisier if it was asked later in the experiment. We tested for variance differences in the same set of streak answers and found no evidence of systematic changes (Appendix Tables H.4 A-C, column 6).

Use of calculation tools. We did not explicitly disallow calculators during either experiment. Many of our questions, however, are such that either an exact answer (as in the case of streak data) or a very good approximation (as in the case of histogram questions where the correct answer puts close to 100% on the middle bin) can easily be derived without one. Nonetheless, in Experiment 2's post-experimental questionnaire, we included a "check all that apply" question to probe the potential use of tools. Roughly half the participants (153 out of 308) reported that they did at least one of the following: (i) wrote down calculations on paper, (ii) used a calculator, or (iii) used an online tool. We compared the reported beliefs of these participants to those who checked none of (i), (ii), or (iii) in the 11-bin and 5-bin histogram questions for the sample size of ten (Appendix Tables H.5.A and H.5.B). We examined these two histograms because for them, the use of calculating tools would help a person with very good statistical understanding to get the right answer (albeit with considerable effort). We find no significant differences. Our interpretation is that calculating tools were likely irrelevant.

Costs of biases and effort exerted. Despite our efforts to sharpen the monetary incentives for accuracy, the actual cost of being biased was modest. We give some numbers here from Experiment 2, where the incentives were stronger. Across all 10 betting questions, participants earned on average $0.77 total, whereas they could have earned $0.83 total in expectation from choosing optimally. From each of the three sets of streak questions (sample sizes of ten, a thousand, and a million), participants earned on average $2.81 rather than $2.86 in expectation from always

[41] We intended to also conduct these order-effect tests in our Experiment 1 data, but could not do so because our software did not record when in the experiment the question was asked.

guessing 50% chance of heads. From each of the three sets of parameter-uncertainty questions, participants earned \$0.90 but could have earned \$0.93 from always guessing 50%. From each histogram question, participants earned in the range \$0.20 to \$0.23 and would have earned roughly an additional \$0.01 from giving the correct responses.[42] Although participants who understood the right answers presumably would not have thrown these pennies away pointlessly, and those who didn't may have had no idea what the cost was, we interpret these results as indicating relatively small financial benefits to participants from improving their answers. For details, see Appendix Table H.6.A and H.6.B.

Anticipating that such monetary measures would not give a very complete account of how engaged they were, we also asked respondents, on a scale from 1 to 5, "Please tell us how much thought and effort you put in to answering the questions on this survey." The response 1 was labeled "I went through as quickly as I could," and 5 was labeled "I put a lot of effort into making the best guesses I could." In Experiment 1, only 3 out of the 76 who answered this question (4%) answered 1, and 48 participants (63%) answered 4 or 5, suggesting that most participants believed that they put at least some effort into their responses. Similarly, in Experiment 2, only 10 out of the 301 who answered this question (3%) answered 1, and 179 participants (59%) answered 4 or 5. We also asked specifically about the histogram questions—"There were a number of questions on which you had to estimate the frequency of each of a set of possible outcomes. How much effort did you exert?"—and in both experiments the distribution of responses was similar to that from the other effort question in both experiments (Appendix Table H.7).

We worried that some of what looks like NBLLN or bin effects may be due to participants lazily putting equal weight on all bins to avoid thinking about or reporting their true beliefs. To test whether the frequency of giving nearly equal probabilities to every category in a histogram was higher among lower-effort participants, we divided responses into "low histogram effort" participants (those who answered 1 or 2) and "high histogram effort" participants (those who answered 3, 4, or 5), and we examined the frequency of giving probabilities in the range 8-10% to every bin in the 11-bin histograms (Appendix G.II.B.b). In Experiment 2, 8% of low-histogram-effort participants gave near-equal probabilities for the sample size of ten, compared with 4% of high-histogram-effort participants ($p = 0.16$). For the sample size of a thousand, the corresponding

[42] The especially weak histogram incentives reflect the flat incentives from the quadratic scoring rule.

numbers are 8% of low-histogram-effort participants compared with 3% of high-histogram-effort participants (*p* = 0.06); and for the sample size of a million, 8% and 2% (*p* = 0.05). In Appendix G.II.B.b, we restrict our analyses of Experiment 2 to individuals who report putting in high effort, and we find that our main qualitative conclusions still hold.

In Experiment 2 (but not Experiment 1), we also asked participants, "When answering the questions, did you try to answer as accurately as you could?" 92% (279 out of 302) responded "Yes."

Within-subject correlation in biases. One of our hypotheses was that participants who exhibited greater GF would also exhibit greater NBLLN. A measure of a participant's degree of GF is the participant-specific slope from a regression of the participants' beliefs about the frequency of a head, given that the first *m* flips were a head, on *m*. The more negative this slope, the greater the GF. 57% of participants in Experiment 1 and 42% in Experiment 2 have a negative slope (Appendix G.I.A.3.d). A measure of a participant's degree of NBLLN for both experiments is the sum of the reported probabilities of "0-5% heads" and "95-100% heads" in the 11-bin histogram for the sample size of 1,000. The correlation between this measure of NBLLN and our regression-based measure of GF is 0.26 (*p* = 0.02) in Experiment 1 and 0.09 (*p* = 0.12) in Experiment 2. We interpret these results as indicating little evidence of a systematic relationship between GF and NBLLN.

8. Discussion and Conclusion

To summarize, we find evidence for the four biases we set out to investigate: GF, NBLLN, ER, and bin effects. Our results replicate key features of prior research, but also point to problems with some existing interpretations. In particular, the LSN interpretation of GF cannot easily accommodate our lack of evidence of backward-looking GF, and existing models of LSN do not predict our finding of GF among randomly drawn outcomes in a sequence. While we replicate Kahneman and Tversky's (1972) results regarding the insensitivity to sample size, we find that—contrary to previous interpretations of this evidence (e.g., Benjamin, Rabin, and Raymond, 2016)—once controlling for bin effects NBLLN does not kick in at sample sizes as small as ten.

In terms of magnitude, the GF we find is weaker than what we anticipated based on results from prior research, while the bin effects are stronger.

Our results make clear that accounting for bin effects is important: It would have been easy to draw mistaken conclusions from the results of some of our histogram questions if interpreted without consideration for bin effects. Growing literatures in economics rely on survey elicitations of people's beliefs about the health consequences of behaviors, such as the likelihood of getting lung cancer from smoking (e.g., Viscusi, 1990), or the distribution of equity returns (e.g., Dominitz and Manski, 2007). These literatures tend to find that people's belief distributions are biased in the direction of being uniform, but bin effects alone could generate that pattern. The training task we used for our histogram-belief elicitations, in which we asked participants to guess the percentage of the population in the United States composed of each of six major racial groups, provides an illustration. While we did not formulate any research questions regarding these data ex ante, participants' responses seem to reflect bin effects. Figures 30 and 31 displays participants' mean estimates in Experiments 1 and 2, respectively, alongside the numbers from the 2010 Census. According to the Census, the most frequent ethnic group is non-Hispanic whites at 63.7% of the population, but participants' mean estimate was only 37% in Experiment 1 and 43% in Experiment 2. The least frequent group is Native Americans, which comprise 0.7% of the population according to the Census but 7% and 4% according to participants in our experiments. If our belief elicitation were taken at face value, it would suggest that people incorrectly believe that whites are a minority in the U.S. and that they overestimate the number of Native Americans by an order of magnitude—but such conclusions are premature until bin effects are taken into account.

The integrated feature of our experimental design made possible one of our central findings, namely that participants' sample beliefs and sequence beliefs cannot be reconciled by a single, internally consistent (even if incorrect) model of the data-generating process to which the rules of probability are applied correctly. Instead, people's beliefs are subject to biases that depend on whether they are asked about the sequence of realizations or the frequencies of outcomes. These findings imply that quasi-Bayesian modeling approaches, which have been the norm in behavioral economics to date, will not be able to predict some important aspects of people's beliefs. These findings also suggest that people's behavior in real-world setting will depend on whether they frame their situation in terms of sequences or distributions of outcomes—which in turn suggests that learning about how people naturally conceptualize the problems they face is a priority for future work.

References

Ahn, David S., and Haluk Ergin (2011). “Framing Contingencies.” *Econometrica*, 78(2): 655–695.

Asparouhova, Elena, Michael Hertzel, and Michael Lemmon (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.

Barberis, Nicholas, Andrei Shleifer, and Robert Vishny (1998). “A model of investor sentiment.” *Journal of Financial Economics*, 49(3): 307–343.

Benartzi, Shlomo, and Richard Thaler (1999). “Risk Aversion or Myopia? Choices in Repeated Gambles and Retirement Investments.” *Management Science*, 45: 364–381.

Benjamin, Daniel J., Matthew Rabin, and Collin Raymond (2016). “A Model of Non-Belief in the Law of Large Numbers.” *Journal of the European Economic Association*, 14: 515–544.

Camerer, Colin F. (1987). “Do Biases in Probability Judgment Matter in Markets?” *American Economic Review*, 77(5): 981–997.

Chen, Daniel, Tobias Moskowitz, and Kelly Shue (2016). “Decision-Making under the Gambler’s Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires” *Quarterly Journal of Economics*, 131(3): 1181–1241.

Croson, Rachel, and James Sundali (2005). “The Gambler’s Fallacy and the Hot Hand: Empirical Data from Casinos.” *Journal of Risk and Uncertainty*, 30, 195–209.

Dominitz, Jeff, and Charles F. Manski (2007). “Expected equity returns and portfolio choice: Evidence from the health and retirement study.” *Journal of the European Economic Association*, *5*(2-3): 369–379.

Fox, Craig R., and Robert T Clemen (2005). “Subjective probability assessment in decision analysis: Partition dependence and bias toward the ignorance prior.” *Management Science*, 51(9): 1417.

Grether, David M. (1980). “Bayes Rule as a Descriptive Model: The Representativeness Heuristic.” *Quarterly Journal of Economics*, 95(3): 537–557.

Haran, Uriel, Don A Moore, and Carey K Morewedge (2010). “A simple remedy for overprecision in judgment.” *Judgment and Decision Making*, 5(7): 467–476.

Kahneman, Daniel, and Amos Tversky (1972). “Subjective Probability: A Judgment of Representativeness.” *Cognitive Psychology*, 3(3): 430–454.

Klos, Alexander, Elke U. Weber, and Martin Weber (2005). "Investment Decisions and Time Horizon: Risk Perception and Risk Behavior in Repeated Gambles." *Management Science*, 51(12): 1777–1790.

McKelvey, Richard, and Talbot Page (1990). "Public and Private Information: An Experimental Study of Information Pooling." *Econometrica, 58*(6): 1321–1339.

Miller, Joshua Benjamin and Adam Sanjurjo (2016). Surprised by the Gambler's and Hot Hand Fallacies? A Truth in the Law of Small Numbers. IGIER Working Paper No. 552. Available at SSRN: https://ssrn.com/abstract=2627354

Oskarsson, An T, Leaf Van Boven, Gary H McClelland, and Reid Hastie (2009). "What's next? Judging sequences of binary events." *Psychological Bulletin*, 135(2): 262–285.

Oppenheimer, Daniel M., and Benoit Monin (2009). "The retrospective gambler's fallacy: Unlikely events, constructing the past, and multiple universes." *Judgment and Decision Making*, 4(5): 326–334.

Peterson, Cameron R., DuCharme, and Ward Edwards (1968). "Sampling Distributions and Probability Revisions." *Journal of Experimental Psychology*, 76 (2), 236–243.

Rabin, Matthew (2002). "Inference by Believers in the Law of Small Numbers." *Quarterly Journal of Economics*, 117(3): 775–816.

Rabin, Matthew, and Dmitri Vayanos (2010). "The Gambler's and Hot-Hand Fallacies: Theory and Applications." *Review of Economic Studies*, 77(2): 730–778.

Rapoport, Amnon, and David V. Budescu (1997). "Randomization in Individual Choice Behavior." *Psychological Review*, 104: 603–617.

Roth, Alvin E., and Michael W. K. Malouf (1979). "Game-Theoretic Models and the Role of Bargaining in Information." *Psychological Review,* 86(6): 574–594.

Sonnemann, Ulrich, Colin F. Camerer, Craig R. Fox, and Thomas Langer (2013). "How psychological framing affects economic market prices in the lab and field." *Proceedings of the National Academy of Sciences*, 110(29): 11779–11784.

Suetens, Sigrid, Claus B. Galbo-Jørgensen, and Jean-Robert Tyran (2016). "Predicting lotto numbers: A natural experiment on the gambler's fallacy and the hot-hand fallacy." *Journal of the European Economic Association*, 14(3): 584–607.

Teigen, Karl Halvor (1974). "Subjective sampling distributions and the additivity of estimates." *Scandinavian Journal of Psychology*, 15: 50–55.

Terrell, Dek (1994). "A Test of the Gambler's Fallacy: Evidence from Pari-mutuel Games." *Journal of Risk and Uncertainty*, 8, 309–317.

Tversky, Amos, and Daniel Kahneman (1971). “Belief in the Law of Small Numbers.” *Psychological Bulletin*, 76: 105–110.

Tversky, Amos, and Daniel Kahneman (1980). “Causal schemas in judgments under uncertainty.” In M. Fishbein (ed.), *Progress in Social Psychology*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc. Reprinted in D. Kahneman, P. Slovic, and A. Tversky (eds.), Judgment *under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press, pp. 117–128.

Tversky, Amos, and Daniel Kahneman (1983). “Extension versus intuitive reasoning: The conjunction fallacy in probability judgment.” *Psychological Review*, 90(4): 293–315.

Tversky, Amos, and Derek J. Koehler (1994). “Support Theory: A Nonextensional Representation of Subjective Probability.” *Psychological Review*, 101: 547–567.

Viscusi, W. Kip (1990). “Do smokers underestimate risks?” *Journal of Political Economy, 98*(6): 1253–1269.

Wheeler, Gloria, and Lee Roy Beach (1968). “Subjective Sampling Distributions and Conservatism.” *Organizational Behavior and Human Performance*, 3(1), 36–46.

Table 1. Number of questions of various types for the two experiments.

Block	Question format	Exp 1	Exp 2
Ten-flip sets	xxHHHTTxHx vs. xxHHHTTxTx with asymmetric payoffs	10	
	"Of the million ten-flip sets, some of them had these first N flips (in order): HHTHH." Bet on H or T next, with asymmetric payoffs		10
	Histograms of 10 flips	3	3
	"Among the sets in which the first X flips came up H, in how many did the X+1th come up H?	10	
	"What percentage of the 10-flip sets have exactly 2 heads and 8 tails?"	10	10
	Streaks in 10 flips: "Some of the sets of 10 flips came up Tails on flips 1, 2, and 3. For all these sets, please estimate the percentage that also came up Tails on the next flip."		10
	Each set has 10 flips in it, numbered 1 through 10. We chose [one, two, five] flip numbers at random. For those sets that came up Heads on those flips, we picked a [second, third, sixth] flip number at random from the remaining [9, 8, 5] flip numbers. What percentage of these also came up Heads?		3
Thousand-flip sets	Histograms of 1000 flips	4	8
	Streaks in 1000 flips		10
	Each set has 1000 flips in it, numbered 1 through 1000. We chose [one, two, five] flip numbers at random. For those sets that came up Heads on those flips, we picked a [second, third, sixth] flip number at random from the remaining [999, 998, 995] flip numbers. What percentage of these also came up Heads?		3
Million-flip sets	Streaks in 1 million flips		10
	Histograms for 1 million flips		3
	Each set has one million flips in it, numbered 1 through 1,000,000. We chose [one, two, five] flip numbers at random. For those sets that came up Heads on those flips, we picked a [second, third, sixth] flip number at random from the remaining [999, 998, 995] flip numbers. What percentage of these also came up Heads?		3
Extra items	Math teacher	4	
	Math quiz	6	6
	Demographics	7	9

Table 2. Estimates of the parameters of the Rabin-Vayanos model of the law of small numbers.

Parameter	(1) Exp 1 *N* = 10	(2) Exp 2 *N* = 10	(3) Exp 2 *N* = 1000	(4) Exp 2 *N* = 1 mill
α	0.031	0.160	0.020	0.047
	(0.022)	(0.031)	(0.021)	(0.022)
δ	0.947	0.621	0.936	0.788
	(0.152)	(0.063)	(0.200)	(0.118)

Note: The parameters are estimated by non-linear least squares from participants' mean beliefs in the streak data, as described in the main text. Standard errors are in parentheses. In each estimation, the number of observations is 10.

Table 3. Percentage of times participants bet on heads.

Number of heads minus tails	(1) Full Sample (Exp 1)	(2) Target- later (Exp 1)	(3) Target- earlier (Exp 1)	(4) Full Sample (Exp 2)
-7				46.7% [15]
-6				66.7% [18]
-5	60.9% [23]	66.7% [12]	54.6% [11]	57.8% [45]
-4				56.9% [102]
-3	59.8% [164]	68.0% [78]	52.3% [86]	62.4% [221]
-2				59.2% [282]
-1	47.6% [313]	47.8% [148]	47.3% [165]	53.7% [505]
0				53.4% [754]
1	49.7% [306]	53.1% [164]	45.8% [142]	46.0% [517]
2				43.6% [291]
3	47.3% [146]	47.9% [71]	46.7% [75]	45.8% [190]
4				50.6% [83]
5	60.0% [35]	53.9% [13]	63.6% [22]	40.5% [42]
6				30.0% [10]
7				40.0% [5]
Total:	51.0% [987]	53.5% [486]	48.5% [501]	51.9% [3080]

Note: Percentage of time participants bet on heads upon observing a sequence with the given difference between the numbers of heads and tails. In Experiment 1, participants were shown five flips. In Experiment 2, the number of flips shown varied from zero to nine, but the difference between heads and tails equal to -9, -8, 8, and 9 had no observations. In both experiments, each participant contributes ten observations. The number of observations from which each percentage is calculated is shown in brackets.

Table 4. Regression of betting-on-heads on realizations of other flips and payoffs.

	(1)	(2)	(3)	(4)	(5)	(6)
	Full Sample	Target-later	Target-earlier	RV Simulation	Full Sample	RV Simulation
	(Exp 1)	(Exp 1)	(Exp 1)	(Exp 1)	(Exp 2)	(Exp 2)
# Heads - # Tails	-0.0095	-0.0199	-0.0006	-0.0175	-0.0300**	-0.0364**
	(0.0076)	(0.0105)	(0.0102)	(0.0102)	(0.0048)	(0.0057)
Target-later sequence	0.0642					
	(0.0352)					
Head option on top	0.0249	0.0120	0.0332			
	(0.0290)	(0.0489)	(0.0463)			
Payoff from betting H						
45	0.3657**	0.4620**	0.3289**		0.2434**	0.4004**
	(0.0515)	(0.0650)	(0.0718)		(0.0220)	(0.0293)
47	0.3702**	0.4349**	0.3730**		0.2790**	0.4471**
	(0.0411)	(0.0561)	(0.0589)		(0.0229)	(0.0305)
49	0.3649**	0.4430**	0.3519**		0.2973**	0.4691**
	(0.0423)	(0.0600)	(0.0564)		(0.0233)	(0.0302)
51	0.5554**	0.6169**	0.5607**		0.7033**	0.5210**
	(0.0513)	(0.0610)	(0.0602)		(0.0247)	(0.0314)
53	0.5431**	0.6313**	0.5223**		0.7819**	0.5751**
	(0.0526)	(0.0636)	(0.0648)		(0.0206)	(0.0292)
55	0.5748**	0.5984**	0.5995**		0.8221**	0.6458**
	(0.0487)	(0.0767)	(0.0463)		(0.0200)	(0.0293)
Estimated heads bias	0.0047	0.0360	-0.0271		0.0003	
	(0.0233)	(0.0330)	(0.0369)		(0.0150)	
# Obs	987	486	501	987	3080	3080

Note: Linear probability models with no constant term. The dependent variable in columns 1, 2, 3, and 5 is a dummy for betting on the heads sequence. Note that in Experiment 2, there was no target-later vs. target-earlier variation, and the betting-on-heads option always appeared on top. As described in the text, columns 4 and 6 display average results from 1000 simulations. Standard errors are clustered by participant for columns 1, 2, 3, and 5, and for columns 4 and 6 the standard errors are adjusted for variance in the coefficients across simulations. "Estimated heads bias" is the predicted value from the regression, setting the first regressor equal to 0, the second and third regressors equal to ½, and the payoff dummies equal to 0, 0, ½, ½, 0 and 0. * $p < 0.05$, ** $p < 0.01$.

Table 5. Comparison of histogram beliefs in sample sizes of *N* = 10 and 1000 with evidence from Kahneman and Tversky (1972).

Sample size of *N* = 10							Sample size of *N* = 1000						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Heads		K&T:	Exp1:	Exp1:	Exp2:	Exp2:			K&T:	Exp1:	Exp1:	Exp2:	Exp2:
	Correct	Medians	Medians	Means	Medians	Means	Heads	Correct	Medians	Medians	Means	Medians	Means
0	0.1%	2.0%	2.0%	6.1%	1.0%	2.2%	≤5%	0.0%	3.0%	5.0%	9.2%	1.0%	2.9%
1	1.0	5.0	5.0	6.4	2.5	3.8	5-15	0.0	5.0	5.0	7.2	3.0	3.9
2	4.4	7.0	8.0	8.0	5.0	5.5	15-25	0.0	7.0	9.0	8.5	5.0	5.2
3	11.7	10.0	10.0	9.0	10.0	9.3	25-35	0.0	10.0	10.0	8.4	9.0	7.8
4	20.5	15.0	10.0	12.3	15.0	15.1	35-45	0.1	15.0	10.0	12.9	13.0	13.5
5	24.6	20.0	14.0	20.0	25.0	28.3	45-55	99.8	21.0	11.0	18.5	28.0	34.2
6	20.5	17.0	10.0	12.7	15.0	15.0	55-65	0.1	15.0	10.0	11.4	13.0	13.3
7	11.7	10.0	10.0	8.9	10.0	9.2	65-75	0.0	10.0	9.0	8.2	9.0	7.8
8	4.4	7.0	6.0	7.3	5.0	5.5	75-85	0.0	5.0	5.0	6.3	5.0	5.2
9	1.0	4.0	5.0	6.5	2.0	3.9	85-95	0.0	5.0	5.0	5.6	2.3	3.9
10	0.1%	1.0%	1.0%	2.8%	1.0%	2.4%	≥95%	0.0%	3.0%	2.5%	4.2%	1.0%	2.5%

Note: "K&T Medians" refers to the numbers eyeballed from Kahneman & Tversky's (1972) Figure 1a.

Figure 1. Belief in likelihood of heads after a streak of heads in sample size of *N* = 10 (Experiment 1).

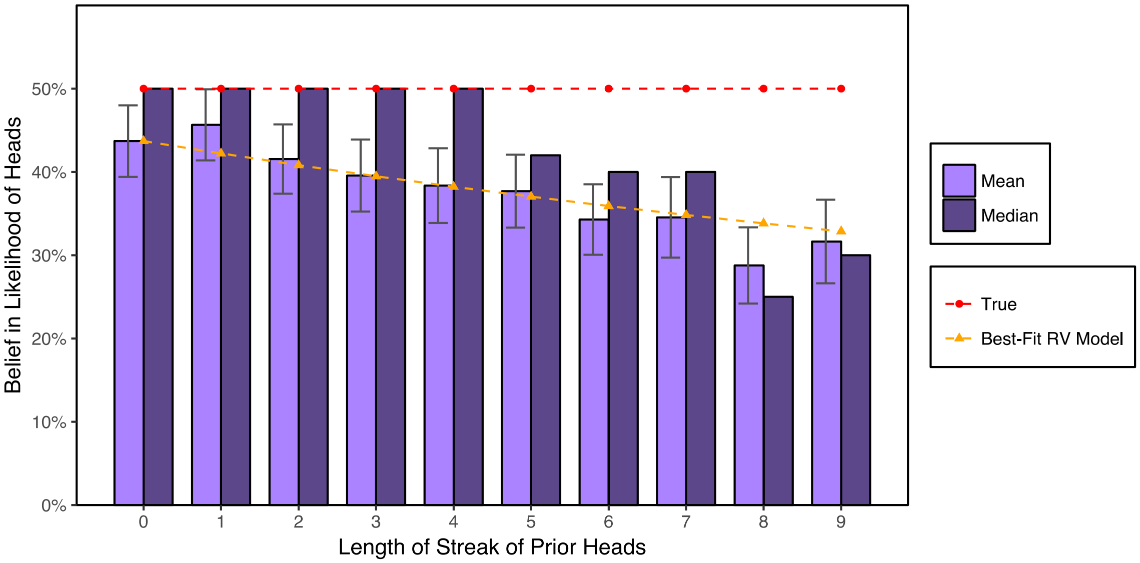


Figure 2. Belief in likelihood of heads after a streak of heads in sample size of *N* = 10, after excluding participants who reported 10% for any streak length (Experiment 1).

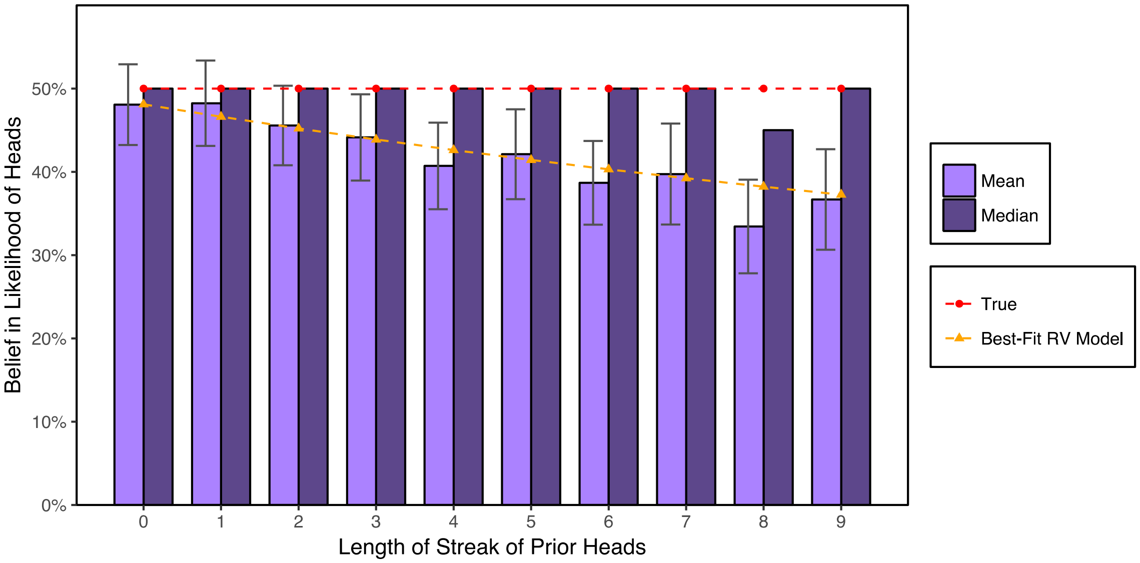


Figure 3. Belief in likelihood of heads after a streak of heads in sample sizes of *N* = 10, 1000, and 1 million (Experiment 2).

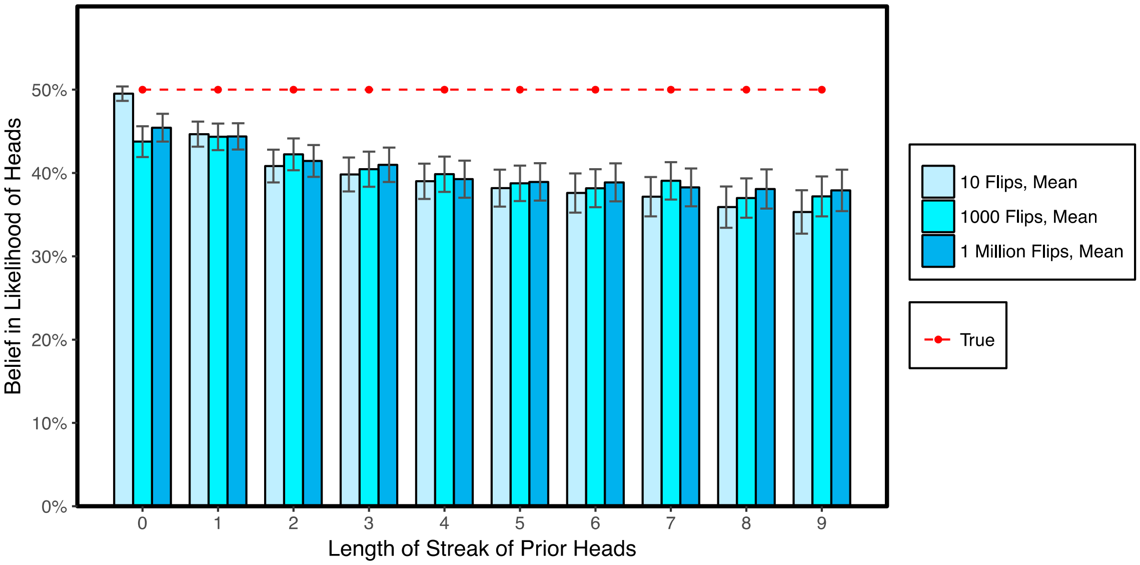


Figure 4. Belief in likelihood of heads after randomly chosen flip locations are known to be heads in sample sizes of *N* = 10, 1000, and 1 million (Experiment 2).

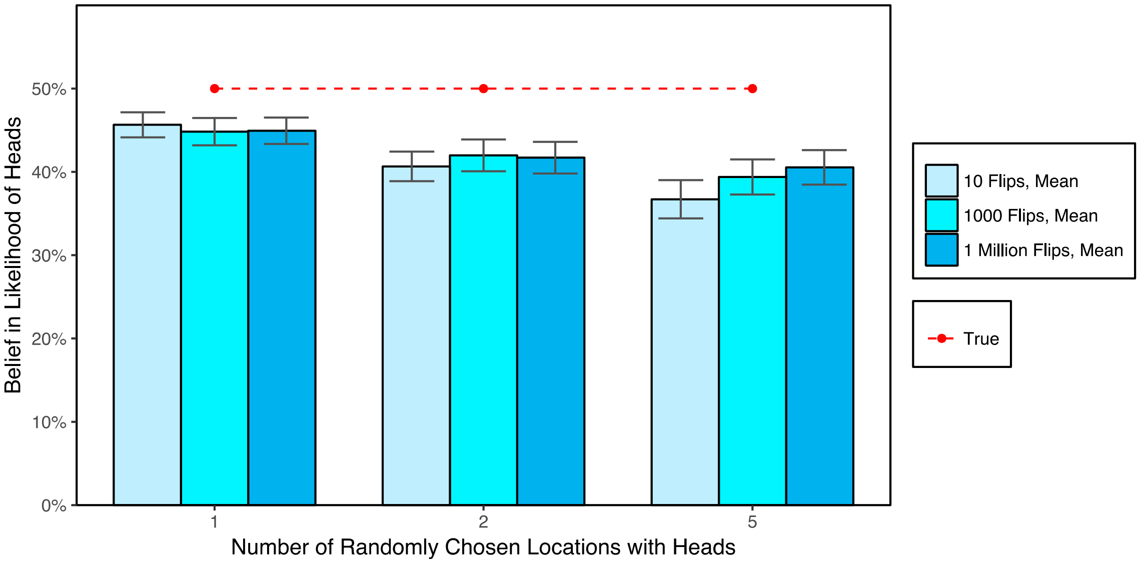


Figure 5. Eleven-bin histogram beliefs in sample size of *N* = 10 (Experiment 1).

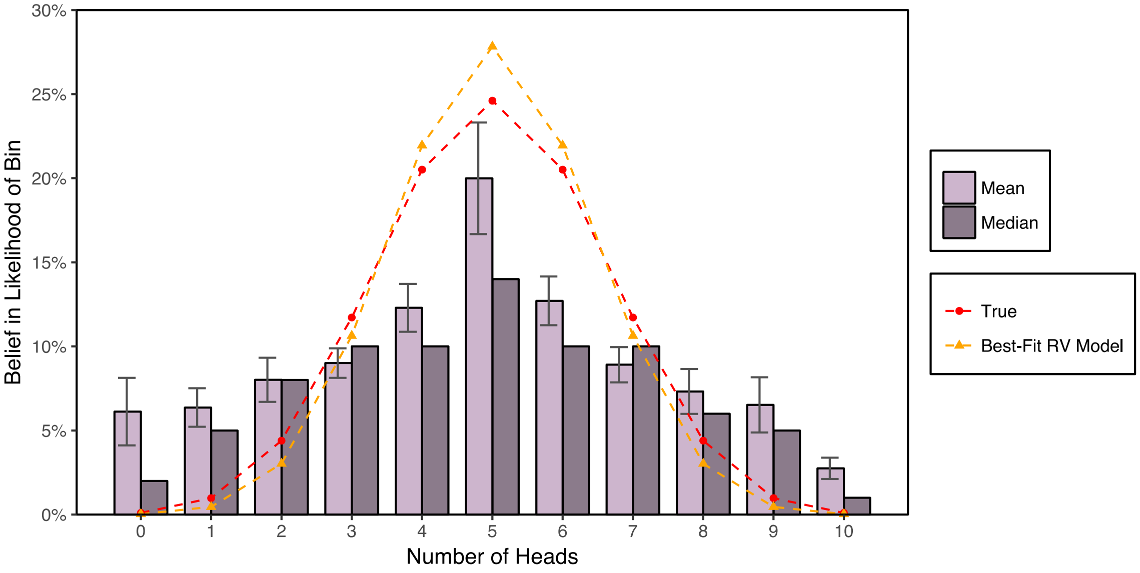


Figure 6. Five-bin histogram beliefs in sample size of *N* = 10 (Experiment 1).

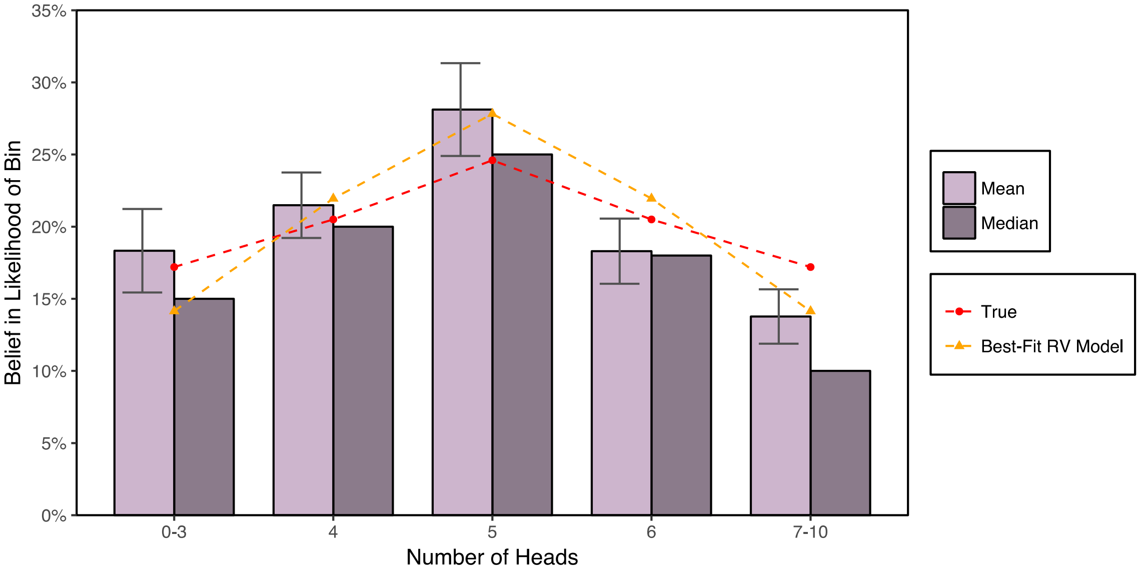


Figure 7. Three-bin histogram beliefs in sample size of *N* = 10 (Experiment 1).

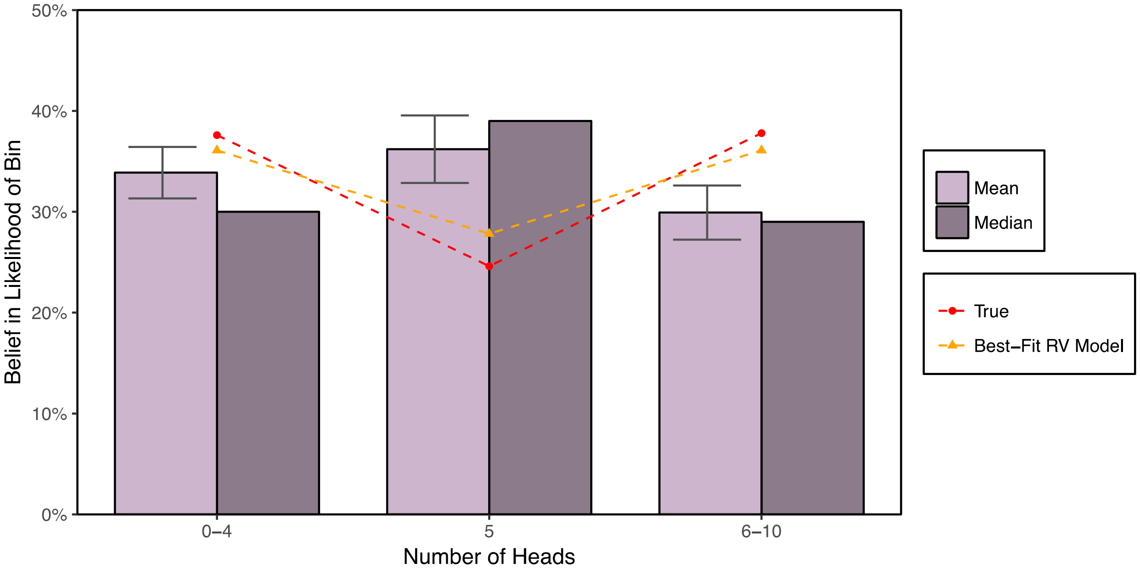


Figure 8. Separate-bin-elicitation histogram beliefs in sample size of *N* = 10 (Experiment 1).

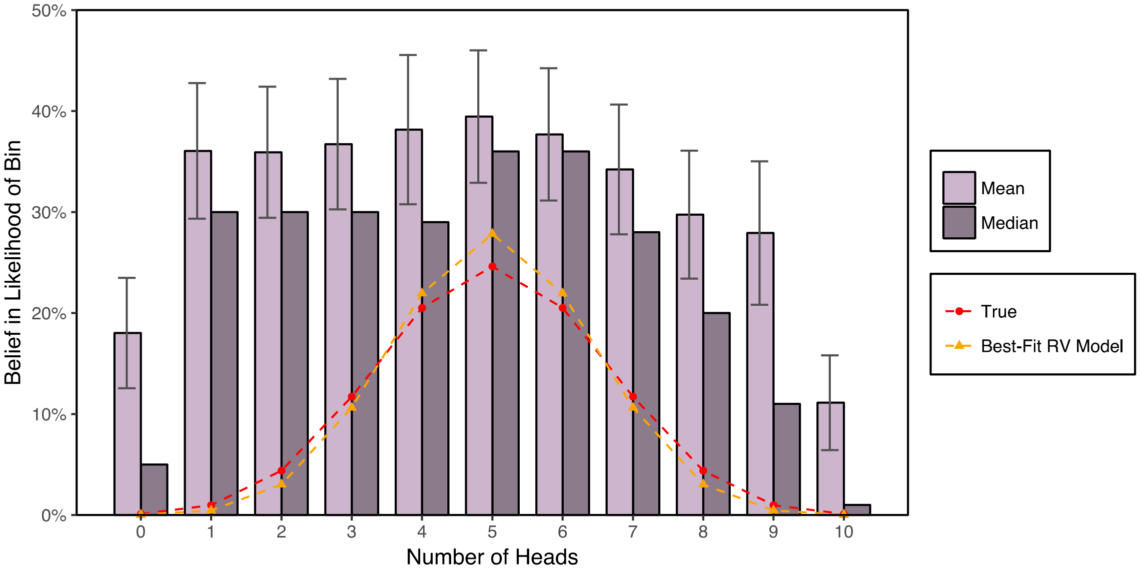


Figure 9. Eleven-bin histogram beliefs in sample size of *N* = 10 (Experiment 2).

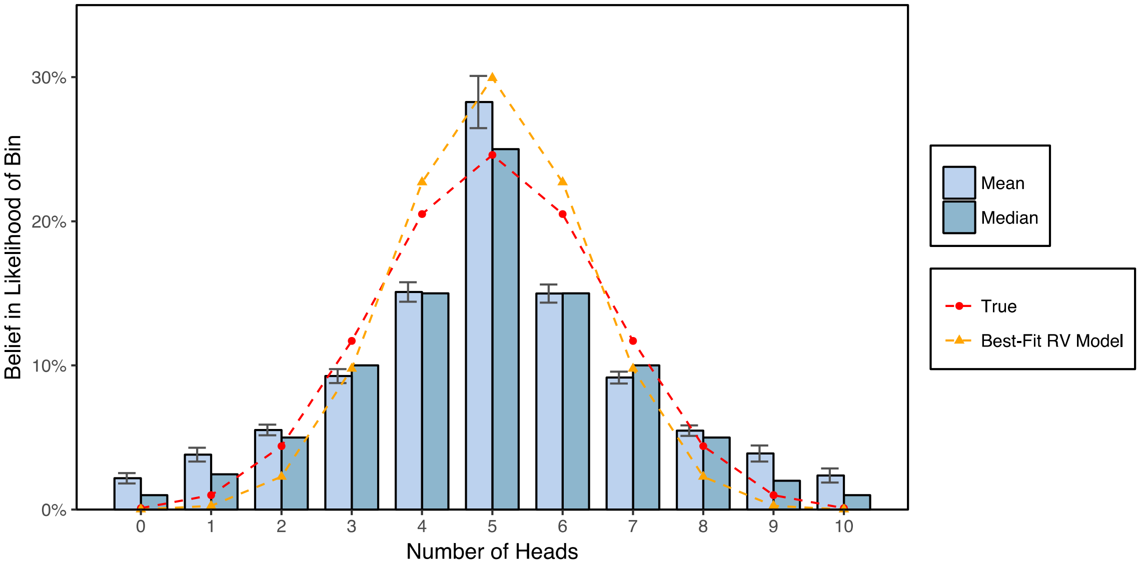


Figure 10. Five-bin histogram beliefs in sample size of *N* = 10 (Experiment 2).

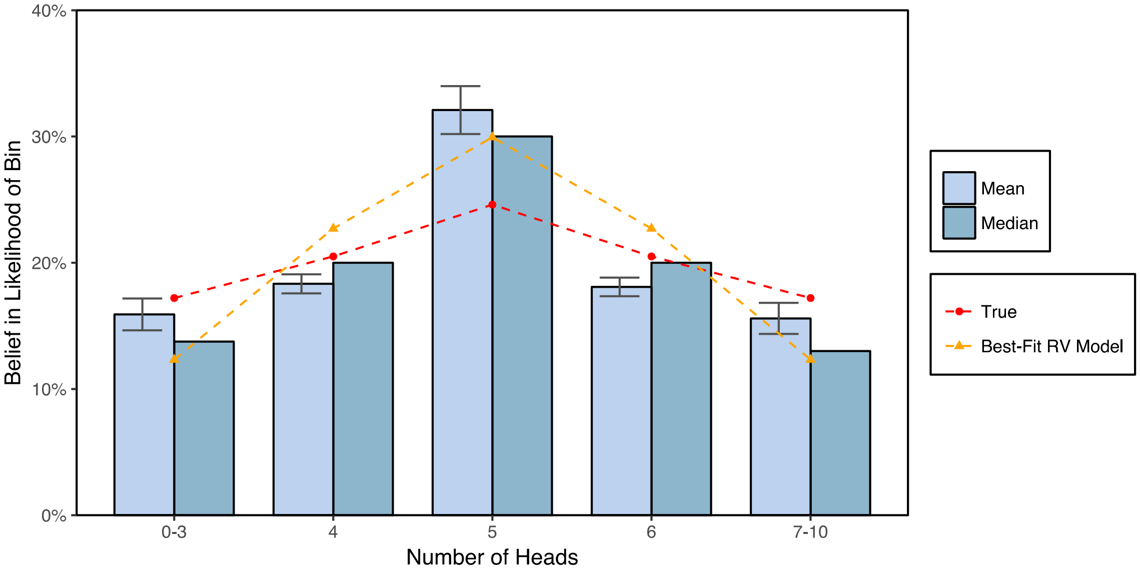


Figure 11. Three-bin histogram beliefs in sample size of *N* = 10 (Experiment 2).

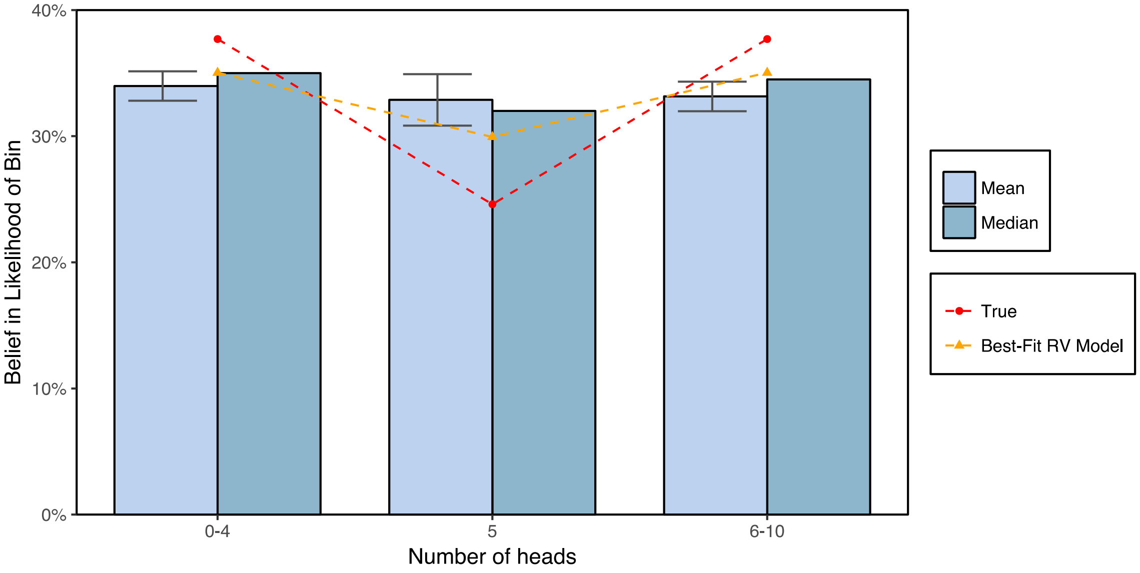


Figure 12. Separate-bin-elicitation histogram beliefs in sample size of *N* = 10 (Experiment 2).

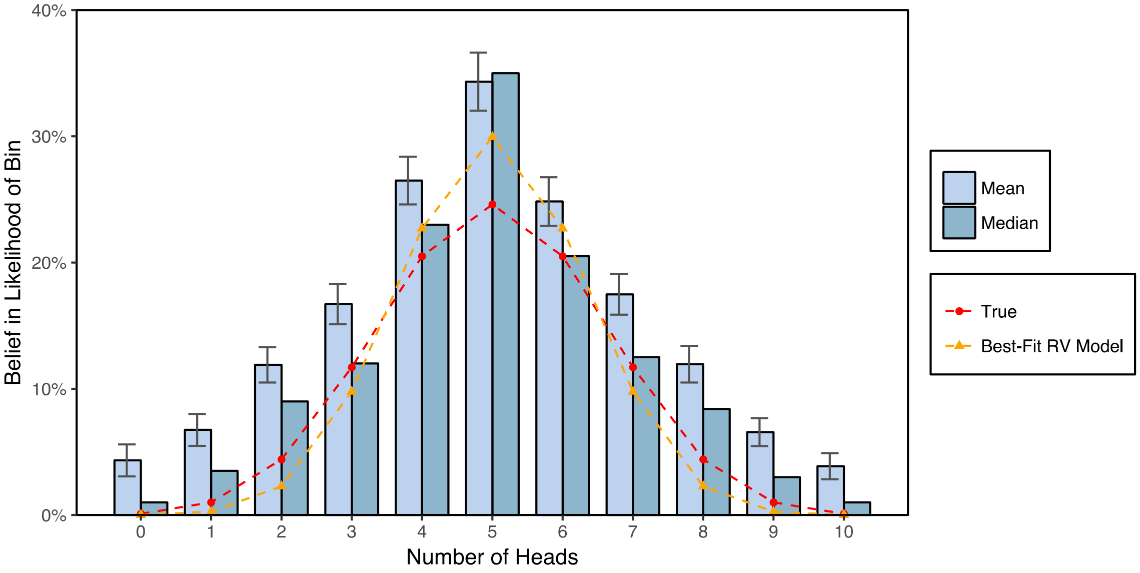


Figure 13. Eleven-bin histogram beliefs in sample size of *N* = 1000 (Experiment 1).

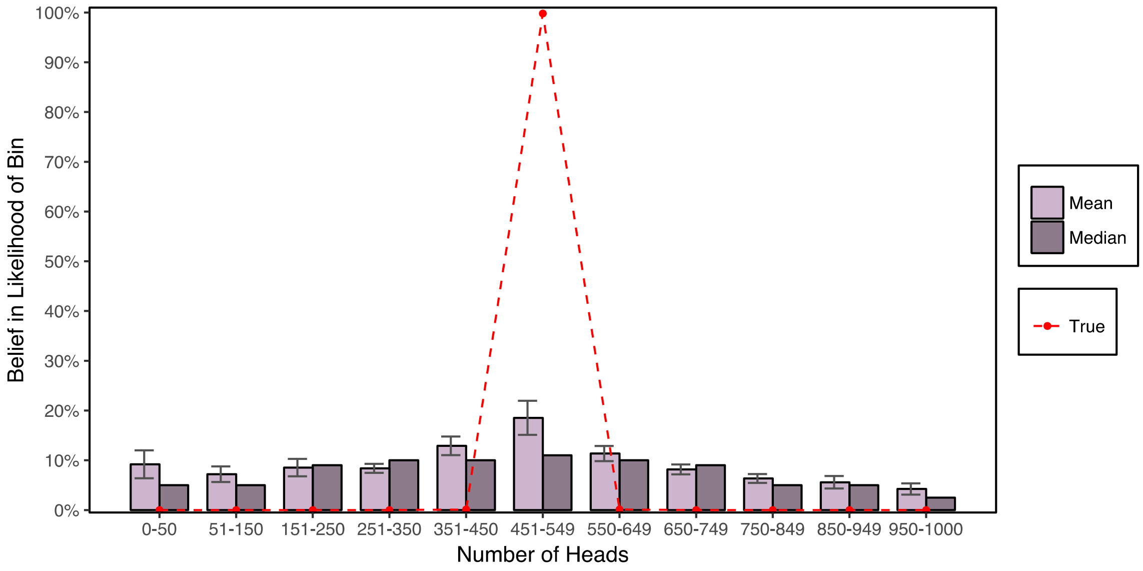


Figure 14. Three-bin histogram beliefs in sample size of *N* = 1000 (Experiment 1).

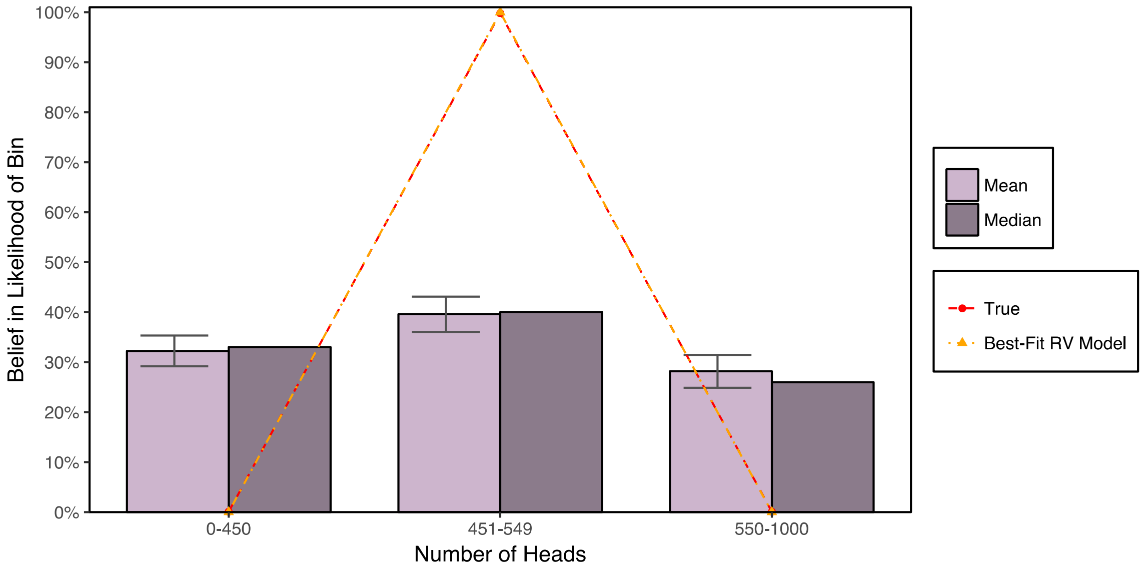


Figure 15. Eleven-bin histogram beliefs in sample size of *N* = 1000 (Experiment 2).

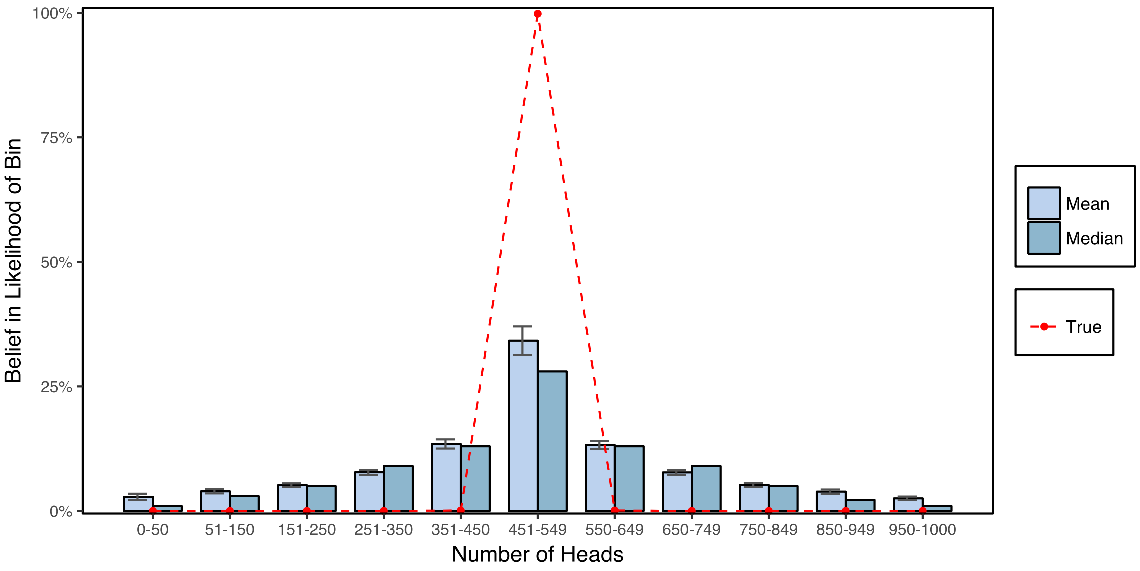


Figure 16. Three-bin histogram beliefs in sample size of *N* = 1000 (Experiment 2).

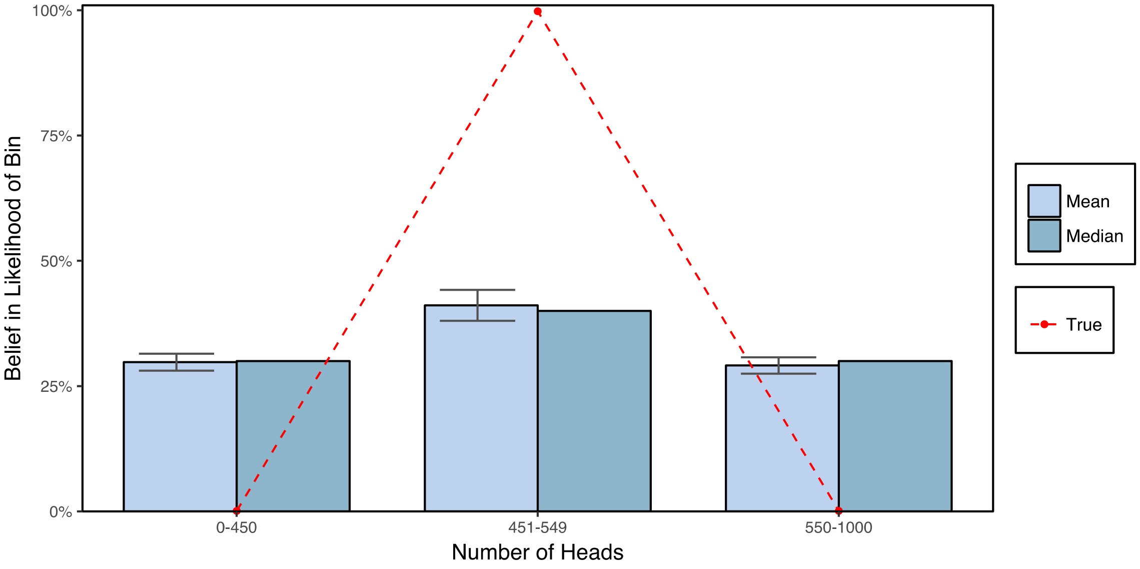


Figure 17. Three-bin histogram beliefs in sample size of *N* = 1000, with the middle bin 481-519 (Experiment 1).

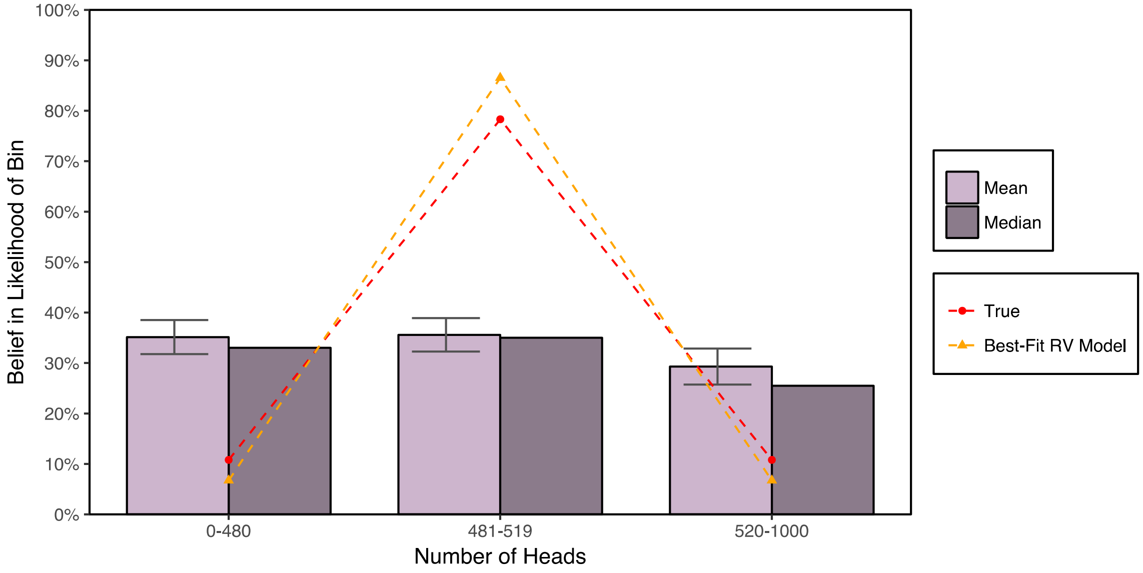


Figure 18. Eleven-bin histogram beliefs in sample size of *N* = 1000, with the middle bin 481-519 (Experiment 2).

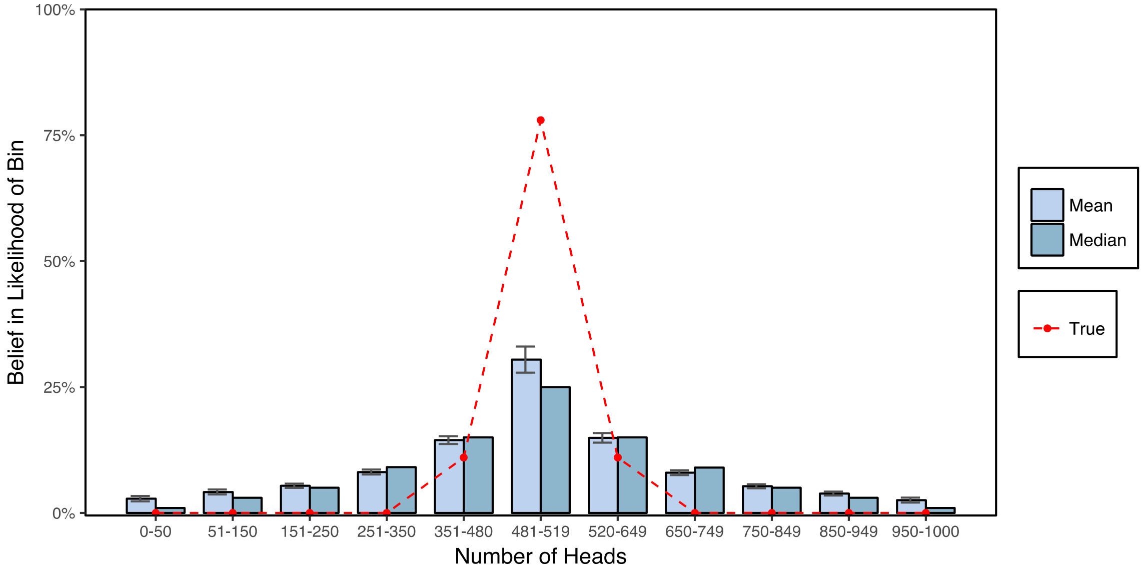


Figure 19. Three-bin histogram beliefs in sample size of *N* = 1000, with the middle bin 481-519 (Experiment 2).

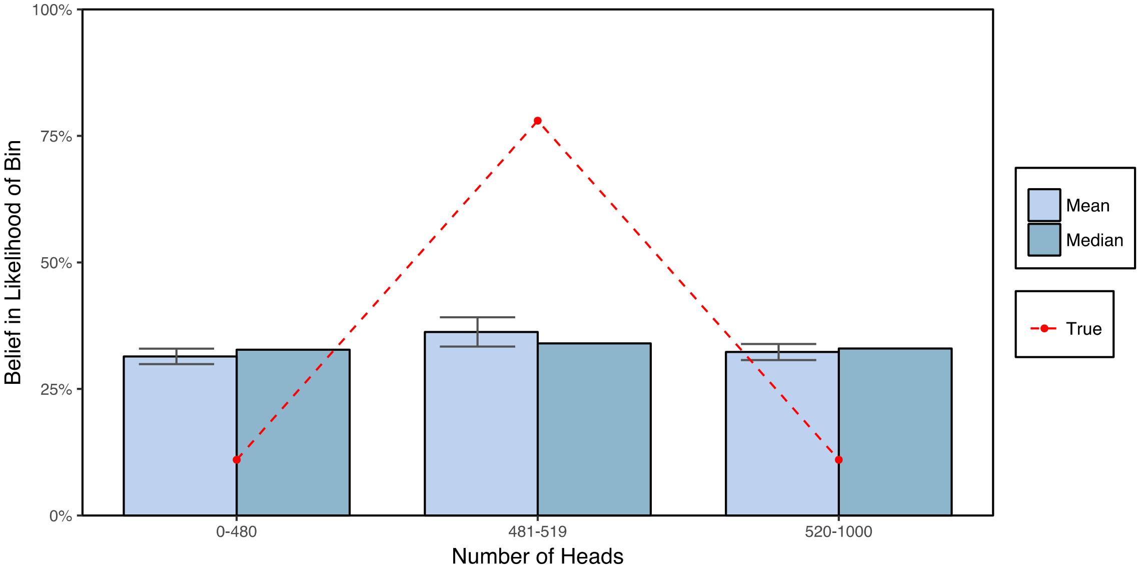


Figure 20. Sample-size neglect for median beliefs in sample sizes of *N* = 10 and 1000 (Experiment 1).

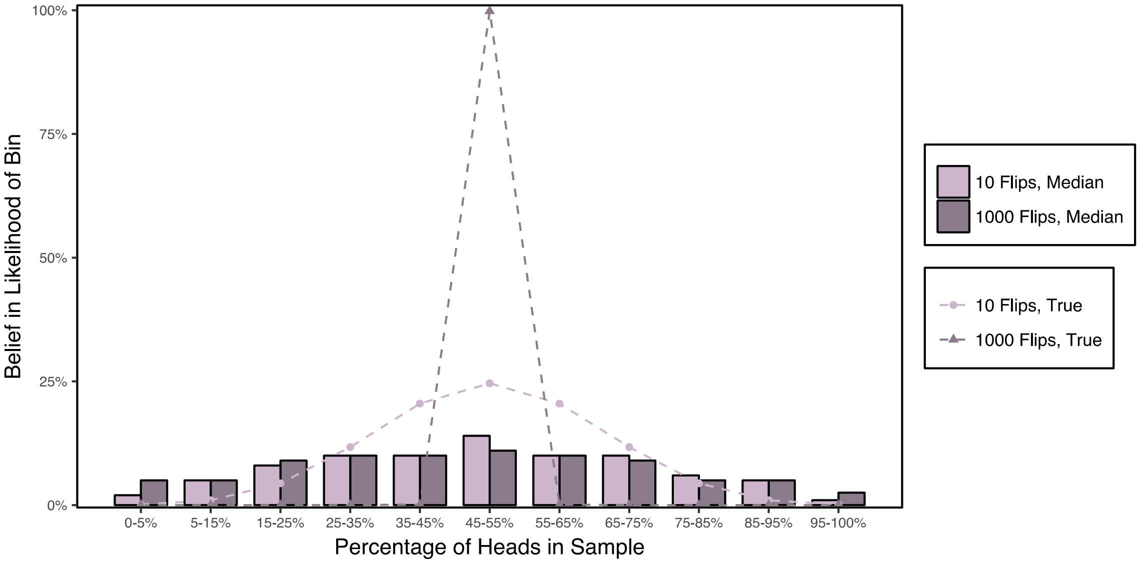


Figure 21. Sample-size neglect for median beliefs in sample sizes of *N* = 10, 1000, and 1 million (Experiment 2).

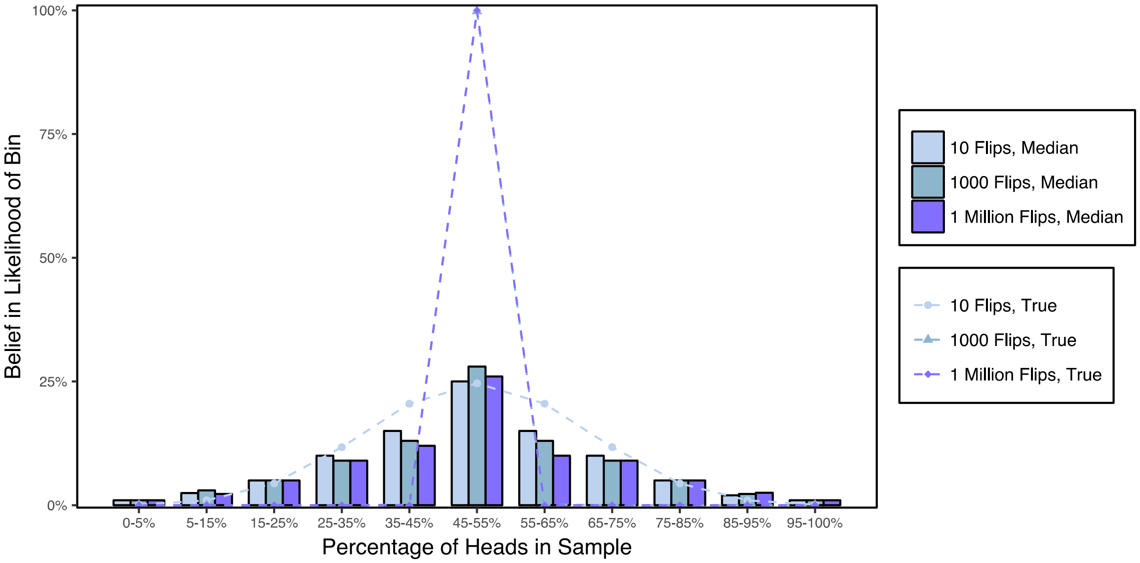


Figure 22. Sample-size neglect for mean beliefs in sample sizes of *N* = 10 and 1000 (Experiment 1).

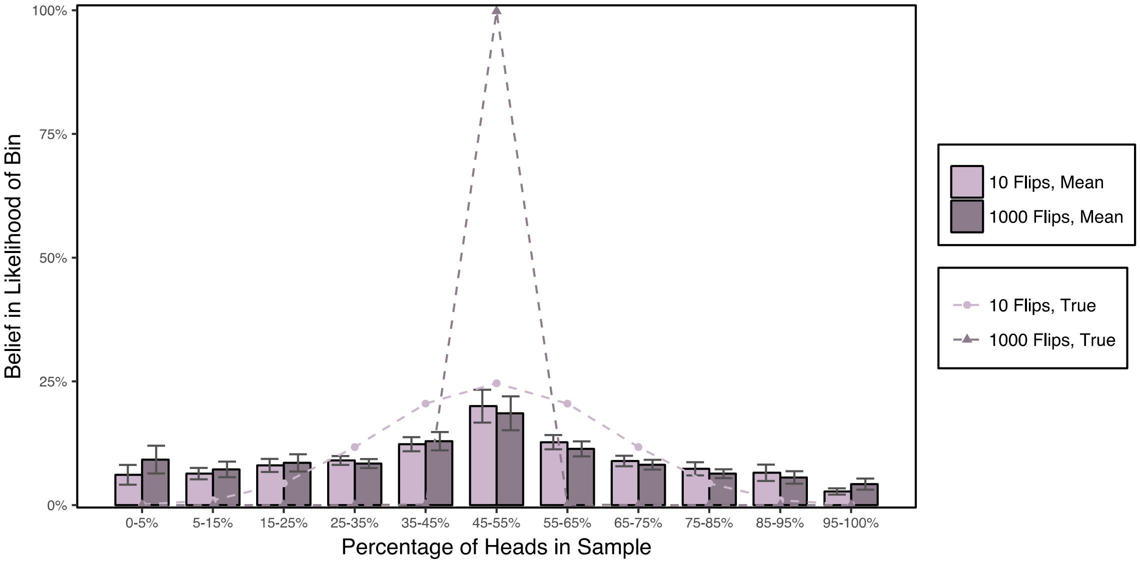


Figure 23. Sample-size neglect for mean beliefs in sample sizes of *N* = 10, 1000, and 1 million (Experiment 2).

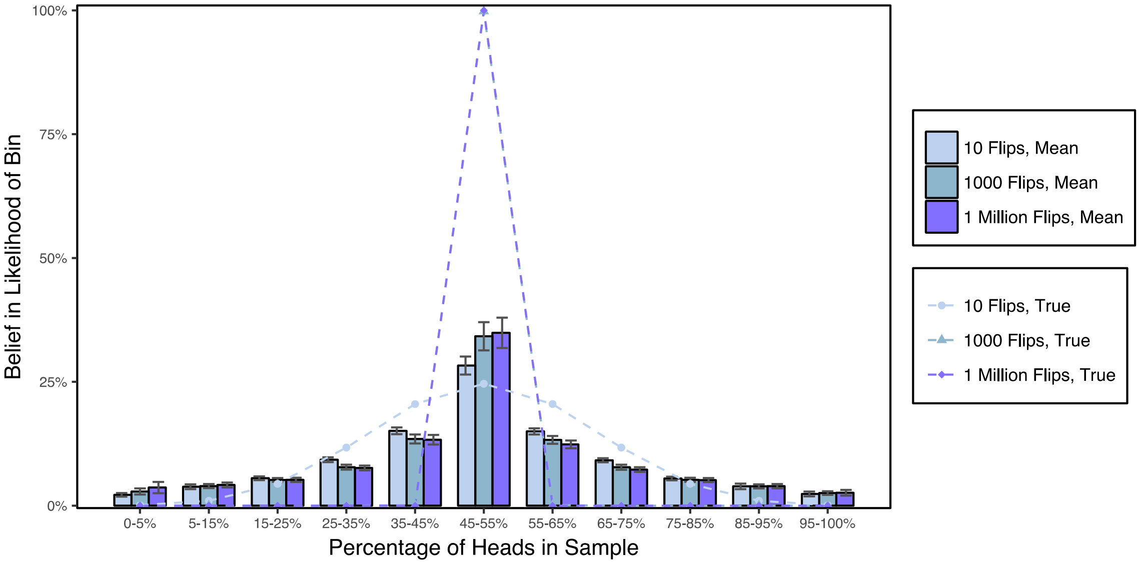


Figure 24. Three-bin histogram beliefs in sample size of *N* = 1000, with the middle bin 500 (Experiment 2).

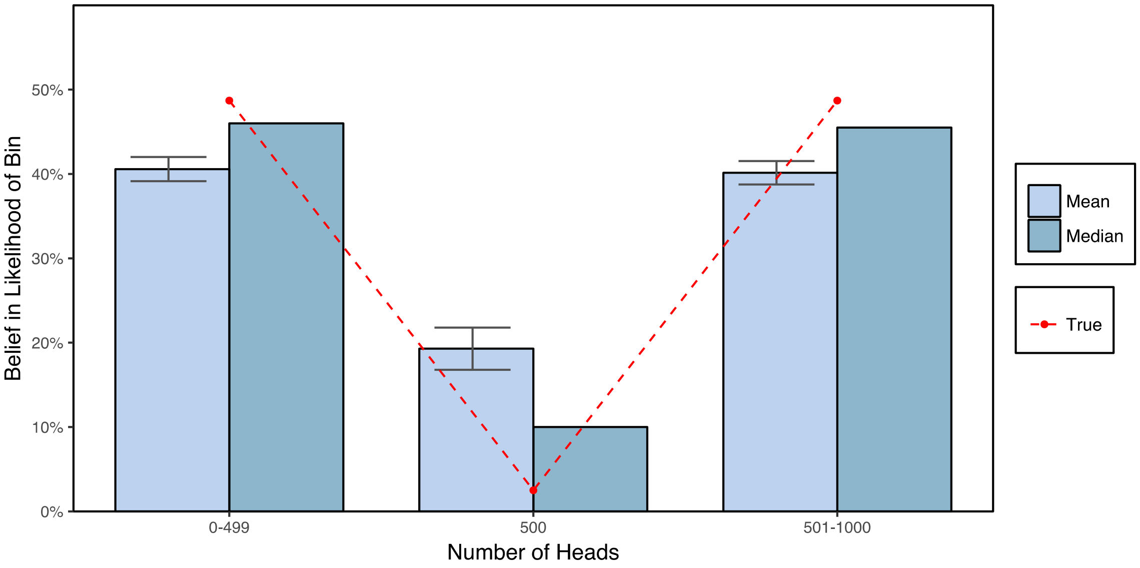


Figure 25. Three-bin histogram beliefs in sample size of *N* = 1000, with the middle bin 494-506 (Experiment 2).

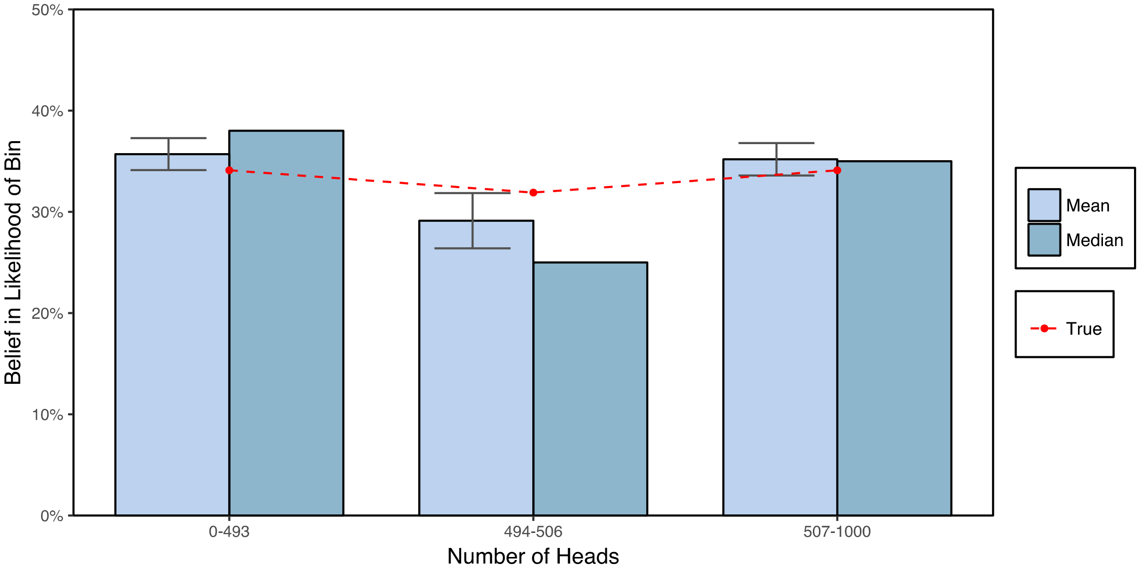


Figure 26. Three-bin histogram beliefs in sample size of *N* = 1000, with the middle bin 910 (Experiment 2).

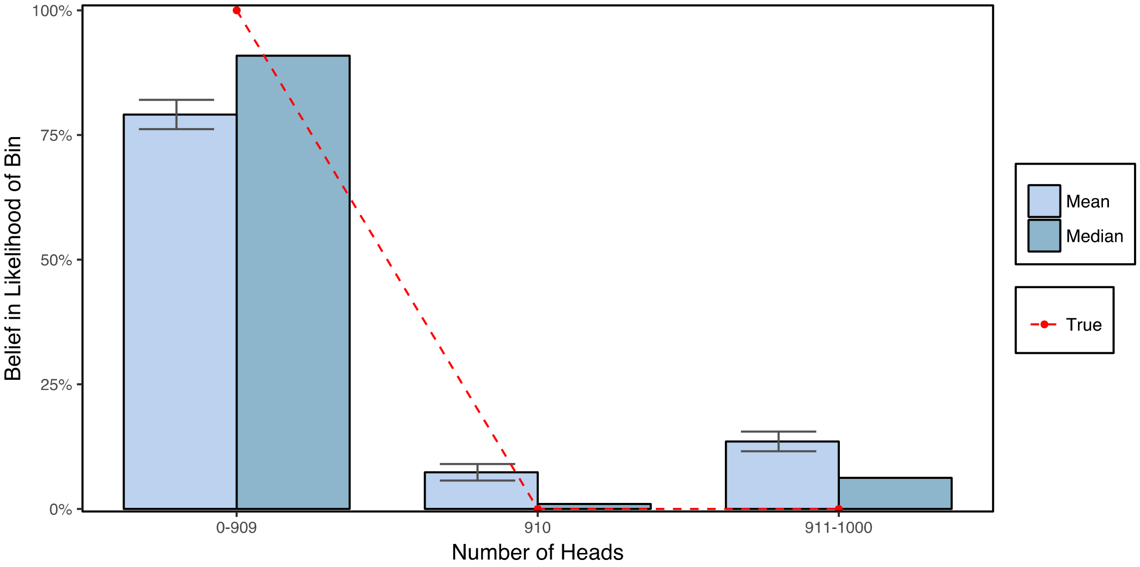


Figure 27. Five-bin histogram beliefs in sample size of *N* = 1000 (Experiment 2).

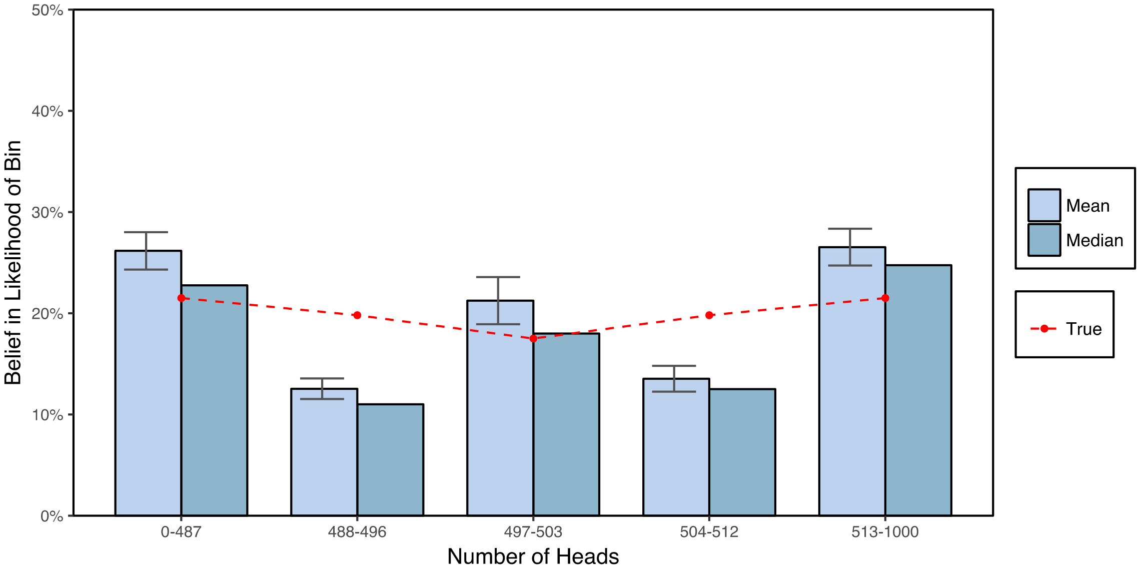


Figure 28. Three-bin histogram beliefs in sample size of *N* = 1 million (Experiment 2).

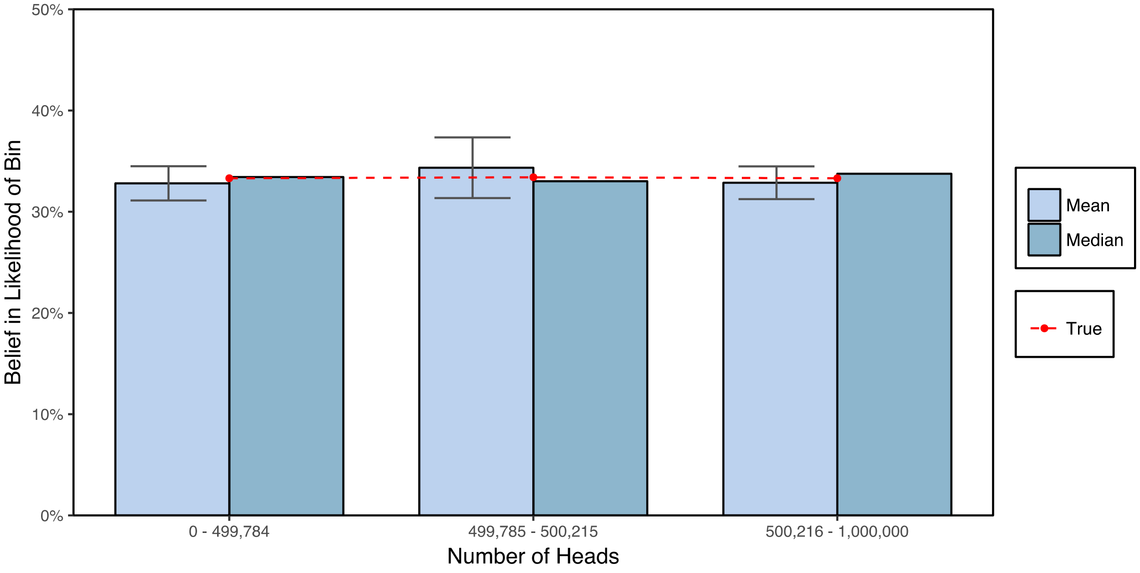


Figure 29. Five-bin histogram beliefs in sample size of *N* = 1 million (Experiment 2).

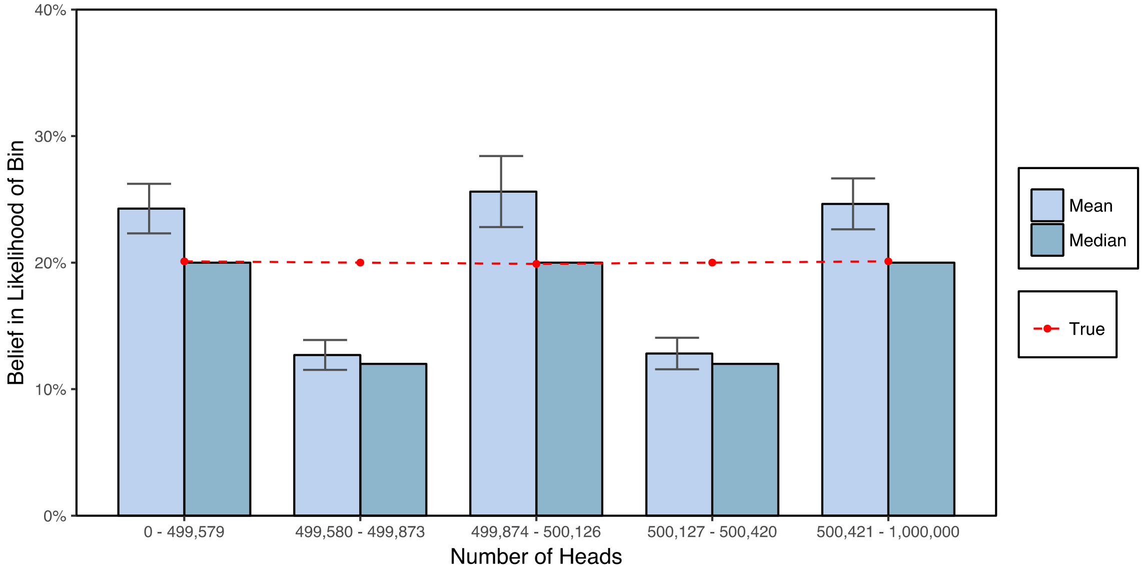


Figure 30. Beliefs about the U.S. ethnic distribution (Experiment 1).

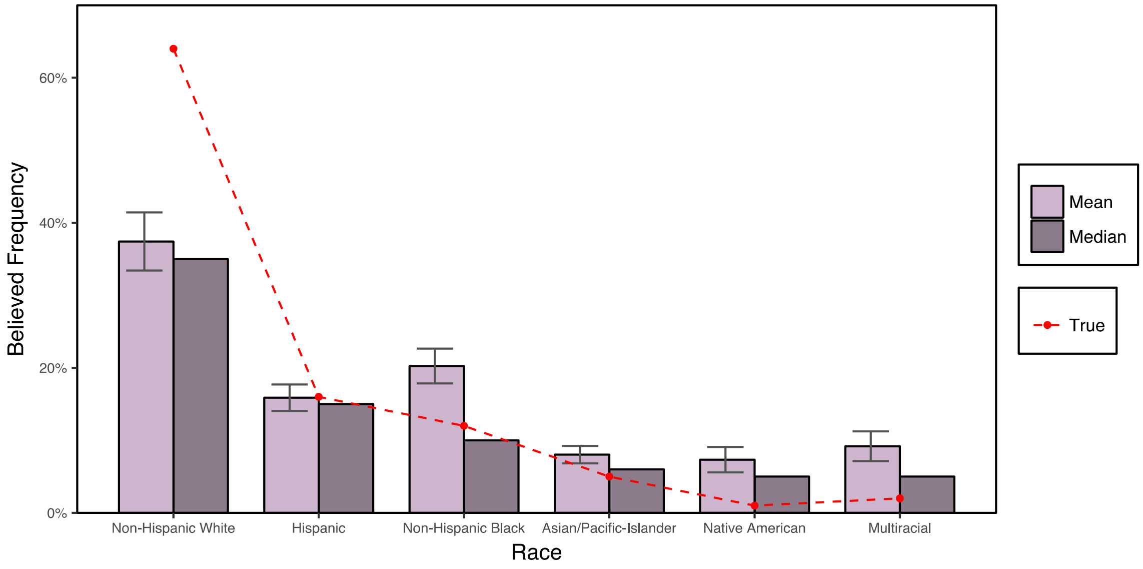
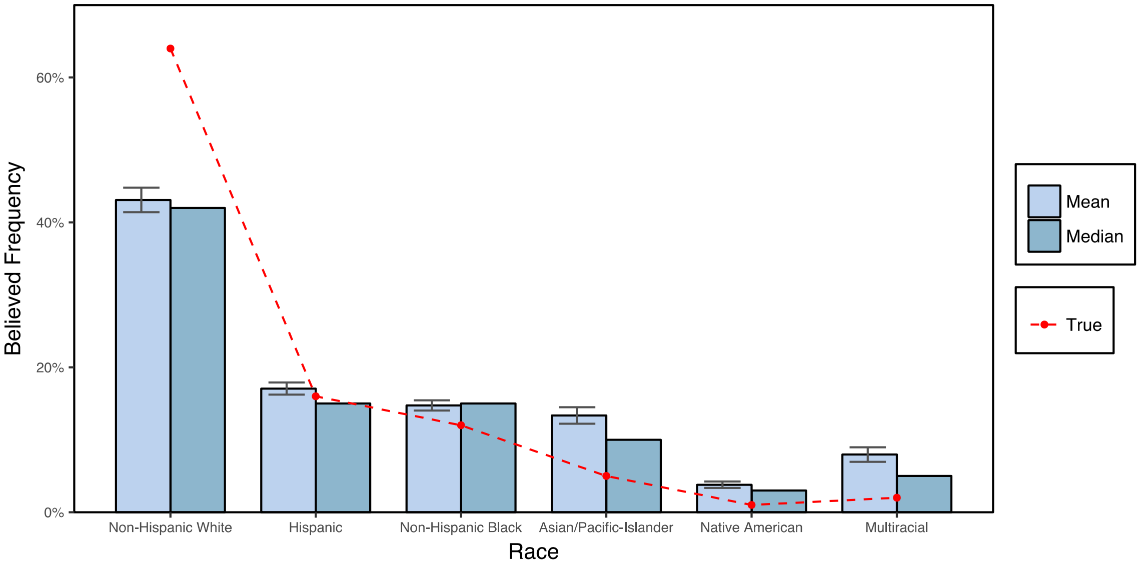


Figure 31. Beliefs about the U.S. ethnic distribution (Experiment 2).


Believed Frequency
Non-Hispanic White
Hispanic
Non-Hispanic Black
Asian/Pacific-Islander
Native American
Multiracial
Race
Mean
Median
True