Misconceptions of Chance: Evidence from an Integrated Experiment

Daniel Benjamin
Cornell University and NBER

Don Moore
University of California—Berkeley

Matthew Rabin*
University of California—Berkeley

This Draft: March 27, 2013

Abstract: This paper describes results of an incentivized experiment investigating biases that jointly imply logically inconsistent beliefs about random samples. We find that people predict reversals of past streaks, consistent with the "gambler's fallacy," but more weakly than in previous experiments and calibrations. Consistent with "support theory," we find that eliciting beliefs about the union of two ranges of outcomes, relative to the two ranges separately, decreases the total probability assigned to those outcomes. When these effects are taken into account, our results suggest that—contrary to previous interpretations—participants' beliefs about sample sizes of 10 are approximately accurate. Yet we strongly confirm earlier findings of "Non-Belief in the Law of Large Numbers" that contradict gambler's fallacy: people vastly exaggerate the likelihood that 1,000-flip samples that would deviate substantially from 50% heads. Because responses to separate questions are logically inconsistent, our experiment rules out a number of alternative, rational interpretations of reported beliefs.

JEL Classification: B49
Keywords: gambler's fallacy, Non-Belief in the Law of Large Numbers, support theory

---

* We are grateful to seminar audiences at Cornell, Harvard, Koc University, Penn, Stanford, UCL, and NYU for helpful comments. We thank Yeon Sik Cho, Tristan Gagnon-Bartsch, Han Jiang, Justin Kang, Muxin Yu, and especially Michael Luo, Alex Rees-Jones, 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 for financial support. E-mail: daniel.benjamin@gmail.com, dmoore@haas.berkeley.edu, rabin@econ.berkeley.edu.

1. Introduction

This paper reports on an incentivized experiment that investigates biases in reasoning about random samples. Our integrated design examines multiple biases concurrently, allowing us to clarify their basic nature and disentangle them from each other and an array of alternative, more rational, interpretations that could be confounding previous experiments.[1]

Although our experiment explored a larger set of related questions and hypotheses, we designed it primarily to study two "sampling biases" and one confounding "elicitation bias."[2] The first sampling bias is a false belief in what Tversky and Kahneman (1971) sardonically dubbed the "Law of Small Numbers" (LSN): people exaggerate how likely it is that small

---

[1] In addition to the more conceptual and methodological motivations for our experiment, we are also motivated on pure replications grounds: despite large literatures related to the biases we study, there are surprisingly few incentivized experiments, and we are unaware of any incentivized, direct evidence on one bias explored here: non-belief in the law of large numbers. See Benjamin, Rabin, and Raymond (2012) for a full review, who find only 6 experiments (from 4 papers) on people's beliefs about sampling distributions—all of them consistent with such non-belief, but none are incentivized. (There are, however, many papers on "conservatism" in belief updating, which we believe is a direct consequence of this same bias, as Kahneman and Tversky (1972) and others have noted.) See Oskarsson, Van Boven, McClelland, and Hastie (2009) and Rabin and Vayanos (2010) for a review of evidence on the law of small numbers.

[2] Although the general topics discussed in the introduction and the next two sections reflect the primary motivations for the experiment and do not contain any ex post hypotheses, they do not include all questions we posed to ourselves. All additional questions are discussed in Appendix B. Because we collected a rich dataset and addressed a number of issues, we were worried about slipping into ex post hypotheses and interpretations during analysis and paper-writing. To prevent this, we wrote a memo to ourselves before we collected any data (final version: February 16, 2010) enumerating our research questions and intended analyses. Although the memo was not originally designed for inclusion in this paper—and hence is not explicit on some matters and not edited to be reader-friendly—we share it verbatim online: http://learnmoore.org/mooredata/Barney/Memo.doc. Our discussion throughout reflects sharper conceptualizations of the issues than we had when we ran the experiment, and some of our statistical tests reflect incremental attempts to understand the data, and were formulated after looking at the data. Specifically: due to shortcomings of our design and weaker-than-anticipated results, the theory-based estimation of the parameters of the Rabin-Vayanos (2010) model was an ex post approach; some of our specific comparisons of bin effects across questions were unplanned; and we did not intend to report results from our training task of estimating the racial composition of the U.S.

samples will reflect the population. 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.

The second sampling bias is an under-appreciation of the effects of sample sizes on the likelihood of different proportions of a sample. The results from Kahneman and Tversky's (1972) (unincentivized) survey, reproduced in Table 1 and Figure 1a, are the best-known evidence: when three groups of participants were asked to estimate distributions for the proportion of heads for 10, 100, and 1000 coin flips, they produced virtually identical estimates regardless of sample size. Even when sample sizes are not fully neglected, people under-appreciate how large random samples will almost surely closely reflect the underlying probabilities. Benjamin, Rabin, and Raymond (2012) formally model the bias and explore its economic implications, calling it "Non-Belief in the Law of Large Numbers (NBLLN)."

Because it is of interest in its own right, and because we believe it is an important caveat to interpreting earlier evidence on NBLLN, we also investigate a feature of belief elicitation: stated beliefs are heavily influenced by the "bins" into which participants are asked to categorize different outcomes. In particular, the more finely a range of outcomes is divided when eliciting beliefs, the more total weight participants assign to the range. For example, Tversky and Koehler (1994) illustrate their closely-related notion of "support theory" by showing that people assign greater total probability to death by cancer, heart disease, and other natural causes (3 bins) than to death by all natural causes (1 bin). This phenomenon—which we call "bin effects" in this context—is of particular importance in NBLLN elicitation: all experiments we know of elicit beliefs in bins where close-to-mean bins have higher probability than far-from-mean bins; therefore, inflated probabilities assigned to the far-from-mean bins could simply be the result of biasing all bins toward equal weight. To our knowledge, no experiments have accounted for this basic confound in NBLLN evidence. Our design varies the binning of outcomes in different questions to address this issue.

In Section 2, we describe our primary research questions and experimental design. We report how we determined our sample size, all data exclusions (if any), all conditions, and all measures in the study. From each of our 104 adult participants, we elicited beliefs in an incentive-

compatible way about the frequency of various outcomes both from flipping a coin 10 times and from flipping it 1000 times. For both of these sample sizes, we generated one million realizations of coin-flip data, and elicited participants' beliefs about the frequency of different sample realizations regarding this *fixed* set of realizations.[3]

We investigate LSN solely for the samples of size 10, but in two different ways. We elicited participants' beliefs about the frequency with which a streak of all heads or all tails will continue, for streaks of length 0 to 9. We separately examined LSN by asking participants to make bets on which of two sequences of coin flips occurs more often; because we vary the sequences we asked about, we can estimate a structural model of how participants believed past—and future—flips affect the likelihood of heads. We examine NBLLN in both samples of size 10 and 1,000 by eliciting participants' beliefs about the frequency distribution over the number of heads. We probe the role of bin effects as a potential confound for NBLLN by eliciting beliefs about outcomes binned in different ways.

We report our main results in Section 3. We find evidence for LSN and GF in both elicitation modes, but the magnitude of the GF appears weaker than it does in previous evidence with which we are familiar. The strongest indication of GF in our data is that the median participant's probability of a heads flip after 9 previous heads was only 32%. Consistent with our conjecture but contrary to existing formal models, we find no evidence of "backward-looking" GF: knowledge of later flip outcomes had no effect on participants' predictions regarding earlier flips in the sequence.[4]

When we elicit participants' histograms using precisely the same categories as Kahneman and Tversky (1972), we find similar results—despite differences in participant populations,

---

[3] As argued by Tversky and Kahneman (1983) and later researchers, posing problems in frequentist vs. probabilistic terms may affect the results. Other features of our design, such as incentive-compatibility, may additionally affect the probabilities that participants report. We did not hypothesize that any of these facets of our design mattered qualitatively and did not test whether they did. When feasible and reasonable, we chose design features favored by researchers who tend to be skeptical of findings of statistical biases.

[4] As far as we are aware, the only existing evidence addressing backward-looking GF is from Oppenheimer and Monin (2009). They find that, when told that a streak occurred at the end of a sample, experimental participants believe that the overall sample was likely to be larger. Oppenheimer and Monin interpret their evidence as supportive of backward-looking GF.

design details, and incentives. If anything, our data show even less belief in the Law of Large Numbers. Their median participant thought that the probability of exactly 5 of 10 babies born in a given hospital being male was 20%, and that the likelihood of between 451 and 549 out of 1,000 babies being male was 21%; we replace babies with coins and find 14% and 12%.

We also find powerful "bin effects," however. When the most likely outcome was one of 3 bins, participants rated it as more likely than when it was one of 11 bins. When an unlikely outcome, such as getting less than 50 tails in 1000 coin flips, received its own bin, participants exaggerated its probability. Participants estimated the likelihood of the bin "less than 50 tails" at 9.2%, yet there was not a single instance of less than 350 tails in our sample of a million 1000-flip sets.

These bin effects call into question some earlier interpretations in the literature. For example, although all data we know of suggest that people exaggerate the likelihood of extreme outcomes in samples of size 10, our results show that this result can be reversed when the bins are changed. When we elicited participants' beliefs with only 3 bins (0-4, 5, and 6-10 heads) of outcomes rather than 11, participants' mean beliefs about the chance of exactly 5 heads from 10 flips changed from 20% to 36%—going from below the correct answer of 25% to above it. Judging from a 5-bin elicitation designed to "neutralize" bin effects, participants actually seem to have approximately accurate beliefs in a sample size of 10 coin flips.

In contrast, our results for samples of 1,000 unambiguously support NBLLN. Participants' responses imply they do not believe that large samples guarantee proportions close to the true base rate. When asked the likelihood of each of the 3 ranges of 0-450, 451-549, and 550-1,000 heads, participants thought that the middle range, which in fact occurs with 99.8%, happened with only 40%, and participants assigned more than 30% to each of the two tails that occur with probability less than .1%. When we elicited beliefs about 0-480, 481-519, and 520-1,000 heads, participants on average thought the 78% middle range occurred with probability 36%, and that each of the two 11% tails occurred with average probability 32%. Although we have no formal test that proves the case fully, based on intuitions from elsewhere and from features of our own data we discuss in Section 3**,** we do not believe it is plausible that bin effects could generate the degree of "compression" of beliefs in these two three-bin elicitations.

Finally, the results from our integrated tests provide evidence about the sources of incorrect beliefs about samples.[5] In particular, because the participants' answers are inconsistent with *any* internally consistent beliefs about the data, we are able to say with some confidence that not *all* of the results come from participants disbelieving or misunderstanding the experimental instructions or thinking that the coin flips were generated from a non-i.i.d. random process.

The inconsistency revealed by our design helps, in turn, address a deeper point about the underlying psychology of the biases we are studying. LSN (and the associated Gambler's Fallacy) can be thought of as one of an array of "quasi-Bayesian" errors in reasoning: people's beliefs accord to a coherent but wrong model of the world.[6] By contrast, we believe that NBLLN is a different sort of error in statistical reasoning. While in principle people might exaggerate the frequency of extreme proportions because of a mistaken belief in positive autocorrelation or because of parameter uncertainty, our findings show that NBLLN is not explained by these "rational" stories. At every opportunity participants had to make specific predictions about sequences, they exhibited variants of false belief in negative autocorrelation along the lines of GF. This is inconsistent with the overly dispersed beliefs in the 1,000-flip questions. We conclude that no single quasi-Bayesian model can explain *both* participants' beliefs about the likelihood of the total number of heads in the sample and their beliefs about specific sequences of flips. While we believe that a quasi-Bayesian interpretation of LSN may capture the correct psychology, we conjecture that NBLLN reflects a failure to appreciate

---

[5] One interpretation proposed for many cognitive biases is "ecological mismatch": while a person's thought process leads to biased beliefs for i.i.d. processes studied in the laboratory, the same thought process would generate appropriate beliefs for the typical, real-world random processes people encounter. However, the coin flips we employ may be the single most ecologically valid experimental paradigm in the history of psychological research. It is likely that every single participant had experience with coin flipping and that all that experience involved flips yielding i.i.d. 50% chance of heads.

[6] Indeed, the quasi-Bayesian approach has been deployed in formal modeling of psychological biases, such as Barberis, Shleifer, and Vishny (1998). Rabin (2002) and Rabin and Vayanos (2010) model LSN as such a "false-model" bias: even in the widespread circumstances where a random process is analogous to drawing balls from an urn *with* replacement, people may perceive it as drawing the balls *without* replacement. Hence people think the "draw" of 4 heads from the 50-50 coin-*cum*-urn means tails are now more likely to be drawn. In this sense, LSN/GF can be thought of as erroneous—but Bayesian—beliefs.

fundamental mathematical rules about how to aggregate the likelihoods of possible sequences into an assessment of the likelihoods of possible proportions in those samples.

We conclude the paper in Section 4 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. We also outline our plans for running another experiment that overcomes some of the limitations of the experiment we report here.

2. Experimental Design and Primary Research Questions

We recruited 104 participants from a busy food court in downtown Pittsburgh, Pennsylvania. This sample size was determined before we had collected data and we did not perform any hypothesis tests until we had finished collecting data. In the final part of the post-experimental questionnaire, participants were asked their gender, age, and annual income (but told that answering was optional); 57% reported female, age ranged from 18 to 69, with a mean of 34 years, and most participants (71%) self-reported income falling in the category "$50,000 to $100,000." The median time to complete the experiment was 27 minutes.

Immediately after their participation, all participants received a choice of either $3 cash or a $5 gift certificate valid at one of the food vendors in the food court. In addition, participants accumulated "lottery tickets" throughout the experiment, and 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, giving a mean probability of winning the prize of about 8%. Winners of this $50 prize were paid by check, sent in the mail. Providing incentive payments in lottery tickets is a standard technique in experiments designed to induce risk-neutrality. As argued by Roth and Malouf (1979), theoretically, a person who maximizes expected utility over money will be risk-neutral over rewards denominated in lottery tickets irrespective of risk aversion over money.

Upon agreeing to participate, participants were taken to a room with several computers at cubicles adjacent to the food court. There they participated in the experiment on a computer provided by the experimenter. The instructions explained that we were interested in participants' beliefs about the likelihood of various coin flips, and it explained how the payment in lottery

tickets would work. We also included an instruction intended to discourage participants from trying to use the questions we asked to infer the correct 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.

All experimental stimulus materials are available online: http://learnmoore.org/mooredata/Barney. After the instructions screen, there were 5 blocks of questions in the experiment, followed by a post-experimental questionnaire. Blocks A, B, C, and D pertain to a million sets of 10 coin flips each, while Block E pertains to a million sets of 1,000 coin flips each. In order to facilitate the flow of the instructions, the computer randomly assigned to each participant an equal chance of facing Blocks A-D before or after Block E. While Block A always appeared before Blocks B, C, and D, the order of Blocks B-D varied randomly between participants.

In the shared general introduction to Blocks A-D, we told participants[7]:

> We flipped a coin ten times. Actually, we had a computer simulate the coin flipping, generating exactly the same type of random series real coins do. This was a fair coin, in the sense that the coin could come up either heads or tails, and there was an equal chance of each. That generated one ten-flip set. Then we did it again, and again, and again, until...we had 1 million ten-flip sets.

In fact, we generated this set of 1 million samples of size 10 after we ran the experiment, using the pseudorandom number generator in Matlab. All questions in Blocks A-D involve eliciting participants' beliefs regarding the frequency of different sequences among this set. In the general introduction to Block E, we told participants:

---

[7] When Block E occurred before Blocks A-D, each of the two instruction screens were the same as described below, except the words "ten" and "thousand" (the sample size in each of the 1 million draws) were interchanged.

> Okay, we really like flipping coins. In addition to the batch of one million ten-flip sets you have just been answering questions about, we also generated a bigger batch of coin flips. This batch also contains a million sets, but each set has a thousand flips.

We also used the pseudorandom number generator in Matlab to create this set of 1 million samples of size 1,000. Each question in Block E elicited a participant's belief about the frequency distribution over the number of heads in this set.

Before we elicited any probabilities—which we always asked about in terms of the proportion of times that an event was realized in the 1 million sets—we truthfully told participants that in each question, the participant would be paid for accuracy according to the quadratic scoring rule:

$$\#\text{ lottery tickets earned} = 100 - 0.01 \times (\text{participant's reported frequency} - \text{actual frequency})^2.$$

Combined with the fact that payments are in lottery tickets, the quadratic scoring rule incentivizes participants to accurately report their expectation of the actual frequency.[8] While the incentives to report *exactly* one's beliefs are quite weak, there is no reason why the flatness of the incentive structure would generate the systematic biases that we aim to examine. In addition to showing participants this formula, the instructions also explained that, because of this payoff structure, "**it pays for you to be as accurate as possible**."

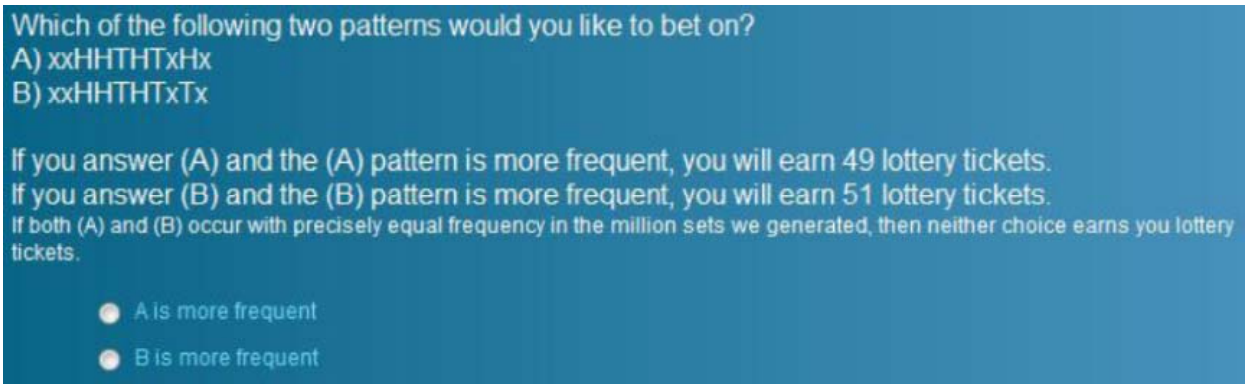
In the remainder of this section, we outline our primary research questions and how the experimental design addresses these questions. In describing the questions we posed to participants during the experiment, we give examples focused on the number of heads. However, for each participant and each block of questions, the computer randomly varied whether the questions were framed in terms of heads or tails.

We test for the presence and magnitude of GF solely in samples of size 10, and we do so in two different ways. First, we elicit participants' beliefs about the frequency with which a streak

[8] As is well-known, the quadratic scoring rule incentivizes a risk-neutral participant to report the mean of her subjective belief distribution about the actual frequency. Hence when we study the average report across participants, we can interpret it as the overall mean belief.

of heads or tails will continue. To grapple far more explicitly than prior work with the full structure of GF, we assess the magnitude of LSN by asking each participant about a streak of each possible length. Specifically, in Block C, participants were asked all 10 possible questions, presented on different screens and in random order, about the conditional probability of a head, conditional on all prior flips being heads. For example, one of the questions was: "Among the sets where the first four flips came up HEADS in what percentage of these did the fifth flip also come up HEADS?" One of the 10 questions was about the first flip: "What proportion of sets have a HEAD as the first flip?"

We also examine GF in another way: we asked participants to make bets about which of two sequences are more common among the 1 million samples of size 10. Both from the instructions and from the questions themselves, we believe that it was clear to participants that the sequences were generated randomly. The randomization of the sequences allows us to examine GF outside the context of extreme samples, such as strings 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. Each of the 10 questions in Block A was a variant of the following example:



The computer randomly assigned one-half probability to the question being of this form, where the 3rd through 7th flip in the sequence are specified, the remainder are unspecified (designated "x"), and the participant must guess whether the sequence is more common with the 9th flip as a head or as a tail. The other half of the time, the question was instead of a 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: does more 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": does more heads occurring *later* in the sequence 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 this.

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:

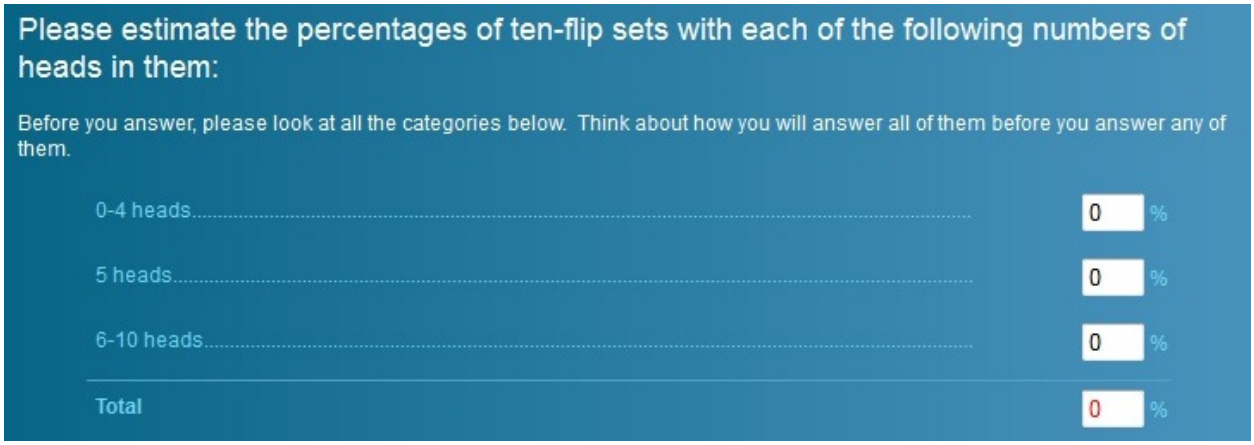
> **Please note: The number of lottery tickets associated with (A) and (B) are chosen randomly between 45 and 55. They do not represent any useful hint toward which pattern is more frequent.**

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. For each participant, 10 of these were selected randomly and independently. This design allowed us to test for both forward-looking and backward-looking GF, as well as to examine beliefs about the likelihood of a head conditional on the total number of known heads (averaging over the sequence combinations).

We test NBLLN by eliciting participants' belief about the frequency distribution of proportions of heads. To give participants experience with the interface for eliciting distributions, an instructions screen required participants to complete a sample question: 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). Participants typed in a

number between 0 and 100 for each group. The screen showed the sum of the percentages, with a sum of 100% required before they could continue to the next screen. Haran, Moore, and Morewedge (2010) called questions of this type, which elicit a histogram of the full subjective probability distribution, SPIES, for Subjective Probability Interval EStimates. We had no a priori hypotheses regarding beliefs about the U.S. ethnic distribution, but we do report the results below.

We elicited SPIES in Block B (and again in D, discussed below) with different bin sizes. Except for Kahneman and Tversky's (1972) evidence on sample sizes of 100 and 1,000, prior evidence on subjective sampling distributions posed the outcomes binned as finely as possible: as 0 heads, 1 head, and so on. In Block B, to assess participants' sensitivity to binning, we elicited SPIES for a sample of size 10 with three questions, which participants faced in a random order. One of the questions asked the participant to estimate, among the 1 million samples of size 10, the frequency of 0-4, 5, and 6-10 heads[9]:



The other two questions also elicited this frequency distribution, but with the events binned differently. The categories for these questions were 0-3, 4, 5, 6, and 7-10 heads and 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 heads. In all three questions, participants could not move on to the next screen until their responses summed to 100%. Participants were told that in each SPIES question, one bin would be randomly selected, and they would be paid for accuracy on that bin using the quadratic scoring rule. While these three different modes were designed to allow us to

[9] Note a bug in the survey format that resulted, for this one question, in half of participants getting response items that asked them to report the percentage of tails but an inconsistent reference to numbers of heads in the question that preceded it. Given that the two interpretations are meaningfully identical, this error does not create a confound.

test for "bin effects," the test for the "fat tails" is comparison of probabilities across the treatments to the correct probabilities. We were interested in whether we replicated the fat tails identified by Kahneman and Tversky (1972) in the eleven-bin treatment and whether the first two variants would yield the same answer.

We elicited SPIES distributions for samples of size 1,000 in Block E. There are four questions in this block, which participants faced in a random order. One question mirrored Kahneman and Tversky's (1972) eleven-bin elicitation of the histogram for a sample of size 1,000, 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 also had 11 categories, and the bold highlights the difference from the first question: 0-50, 51-150, 151-250, 251-350, 351-**480, 481-519, 520**-649, 650-749, 750-849, 850-949, and 950-1000. A third and fourth each had 3 categories: 0-450, 451-549, 550-1000; and 0-480, 481-519, 520-1000. The survey reminded participants that in each question, one bin would be randomly selected, and they would be paid for accuracy on that bin using the quadratic scoring rule. The first of these variants corresponds to Kahneman and Tversky (1972), and we investigated whether we would find similar results. Each of the four treatments allows us to test easily for overly fat tails: the probability of 451-549 heads is over 99%, and the probability of 481-519 is 78%.

We investigate the presence of bin effects by comparing the frequency distributions of outcomes across the different bin structures. Support theory predicts that collapsing any two ranges would decrease total weight on those ranges and enhance the total probability on alternatives that remain fixed. So, for instance, the likelihood assigned to the range 451-549 is predicted to be greater when we elicited it with only two other ranges than when we elicited it with ten other ranges.

In Block D we investigated bin effects in another way. We asked questions of the form: "What percentage of ten-flip sets include exactly 4 HEADS and 6 TAILS?" There were 11 separate questions, on separate screens presented in a random order, separating out each of the possible outcomes for a sample of size 10. Support theory predicts that the total probability assigned to the event "not 4 heads" is will be lower when the alternative is described in terms of

10 bins (0, 1, 2, 3, 5, 6, 7, 8, 9, and 10 heads) rather than one bin (0-3 and 5-10 heads).[10] More simply, Block D allowed us to test whether the probability assigned to a particular outcome, such as 4 heads, would be greater than the probability assigned to the same outcome when beliefs were elicited as a histogram over each possible outcome.

Finally, we intended to investigate a prediction based on the notion of "representativeness" that interplays with LSN, NBLLN, and support theory. In the context of inference, Camerer (1987) described a hypothesis, "exact representativeness," according to which people judge a state of the world to be more likely than its true probability if the observed sample exactly matches the sample that would be most representative of the state. Applied to our context of subjective sampling distributions, an analogous hypothesis would be that a sample of exactly 50% heads is judged to be more likely than its true probability. We can test this hypothesis in two ways. First, we can examine our 5-bin elicitation for a sample of size 10: 0-3, 4, 5, 6, and 7-10 heads. While bin effects predicts that 4, 5, and 6 heads taken together will be assigned probabilities that are too low, the exact representativeness hypothesis predicts that participants will assign relatively excessive probability to 5 heads.[11] Second, we can compare the two 3-bin elicitations for a sample of size 1,000: 0-450, 451-549, 550-1000; and 0-480, 481-519, 520-1000. Because NBLLN would cause people to expect the tails to be far fatter than they actually are, we

---

[10] An implication of this prediction is that, when three or more bins are elicited, the sum of participants' probabilities will exceed 100% when each probability was elicited separately. Let $p(x|P)$ be the reported probability for event *x* given partition of state space *P*. Then we should expect $p(x|\{x\},\{y.z\}) + p(y|\{x\},\{y.z\}) + p(z|\{x\},\{y.z\}) > p(x|\{x,y.z\}) + p(y|\{x,y.z\}) + p(z|\{x,y.z\})$. We are aware of one previous test of whether participants' 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. We caution, however, that the belief elicitation format in Block D may not have isolated bin effects as cleanly as the belief elicitation format in Blocks B and E, because Block D may focus 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. That is, roughly, asking only $p(x|\{x\},\{y.z\})$ may yield a higher answer for $p(x|\{x\},\{y.z\})$ than if we asked both $p(x|\{x\},\{y.z\})$ and $p(\{y.z\}|\{x\},\{y.z\})$.

[11] In our protocol document (see online supplement), we mistakenly intended to examine the 3-bin elicitation: 0-4, 5, and 6-10 heads. In that elicitation, however, the probability of 5 heads is lower than the probabilities of the other events. Hence both bin effects and exact representativeness would lead to exaggerating the probability of 5 heads, so the test for exact representativeness is confounded.

expected that both 481-519 heads and 451-549 heads would be judged to be far less likely than they actually are. However, since 481-519 heads is more similar to exactly 50% heads than 451-549 heads is, the exact representativeness hypothesis would be strongly supported if the former were judged to be more likely than the latter.[12]

A major motivation for our experiment was to assess whether GF and NBLLN would be exhibited within the same participants in an internally inconsistent way—causing participants to implicitly hold different beliefs about the same event depending on whether we elicited sequence beliefs or sample beliefs. While we can address this question by estimating parametric models of GF and NBLLN, we also intended to compare participants' beliefs about the relative likelihood of 9 heads / 1 tail samples versus 10 head samples with participants' beliefs about the likelihood of a tail following 9 heads.

3. Results

We begin with evidence on GF, the most easily isolated bias, as reflected in participants' beliefs about likely sequences. We then turn to beliefs about sample proportions, starting with the bin effects, and then, in light of the bin effects, we interpret the evidence on NBLLN. Finally, we tie discussion of the biases together, arguing why the histograms participants proposed for distributions reflect a NBLLN that is separate from and inconsistent with their beliefs about likely sequences of coin flips.

We interpret our data as, overall, clearly supportive of GF, but certain details of the evidence suggest that existing models of the bias should be modified and that many participants were confused by our elicitation techniques in ways that seem orthogonal to interesting biases. The most straightforward evidence comes from eliciting the conditional probability of a head following a string of heads. These data are displayed in Figure 5a. The gambler's fallacy predicts that participants' beliefs about the frequency of a head on the m^{th} flip, given that the first

[12] Such a finding could also be understood as a "conjunction fallacy" (Tversky and Kahneman, 1983) in which one event is judged more likely than a second because the first event seems more representative of the state of the world, even though the second event implies the first.

m-1 flips were all heads, will be decreasing in *m*, and that all values beyond the first will be less than the correct one of 50%. In fact, participants' average judgment of the probability of a head on the *first* flip was 44%, significantly below 50%. Recalling that these and other reports of "heads" were really balanced both heads and tails, this answer is incoherent and apparently due to a comprehension problem we discuss below. Moreover, on average participants judged the probability of a head after 1 head to be 46%, insignificantly higher than their reported probability of a head on the first flip ($p = .44$). Participants' beliefs appear to be consistent with the gambler's fallacy, however, for $m \geq 2$. Average judged frequency of a head following 9 heads was 32%, significantly below ($p < .01$) the mean belief of 44% following one head. (Here and hereafter, unless stated otherwise, all *p*-values are from two-tailed, paired *t*-tests.) The mean values are monotonically decreasing over this range, and a linear regression of participants' judged frequency on *m* yields a coefficient of -1.7 percentage points with a standard error of 0.2.

The fact that the mean frequency was less than 50% for the first head is hard to reconcile with any psychological model. Closer investigation of the data reveals that many participants reported a frequency of 10% for some values of *m*, suggesting that these participants were confused about what the question was asking. Figure 5b displays the data after dropping all participants who reported 10% for some value of *m*, leaving 69 participants in the sample. Now the mean reported probability of a head on the first flip is 48% with a standard error of 2.5% and hence is not statistically distinguishable from 50%. On average participants judged the probability of a head after 1 head to be essentially the same as the unconditional probability. As in the unadjusted data above, however, participants' beliefs appear to be consistent with the gambler's fallacy for $m \geq 2$, and the slope of the regression of reported frequency on *m* is similar.[13]

The second kind of evidence about the gambler's fallacy comes from participants' bets between pairs of 10-flip sequences. Table 2 displays the fraction of times participants bet on the

[13] At the individual level, the gambler's fallacy appears to be the predominant pattern of belief. We confirmed this in several ways, including individual-level regressions. We report here the simplest and crudest way of categorizing subjects as "pure types": of the 69 participants who did *not* answer 10% for any of these questions, 19 always gave the correct answer of 50%; of the remaining 50 subjects, 13 reported all probabilities at least weakly less than 50%, as implied by the gambler's fallacy, and only 1 reported all probabilities at least weakly greater than 50%, inconsistent with the gambler's fallacy.

heads option as a function of the number of heads in the five known flips. 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).

We turn to regression analysis to control for some of the sources of noise. Table 3 shows linear probability models where the dependent variable is a dummy for betting on the target flip being a head. In column 1, the independent variables are five dummies, one for each of the five flips closest to the target flip being a head. One coefficient is statistically significant: if the 2nd-closest flip is a head, the respondent is 7 percentage points less likely to bet on the heads sequence. However, since the standard errors for the effects of the flip realizations are fairly large, we cannot draw strong inferences from these data. For example, we can statistically distinguish the coefficient on the 2nd-closest flip only from that of the 5th-closest flip ($p < .01$) but not the others.[14] Nonetheless, because of the 2nd-closest flip effect and because the point estimates are negative for 4 of the 5 flip dummies (the exception being the 5th-closest flip), we interpret the evidence in column 1 as broadly consistent with the gambler's fallacy. Column 2 includes several controls—a dummy for whether the sequence was target-later (vs target-earlier), a dummy for whether the betting-on-heads option appeared above (vs below) the betting-on-tails option, and dummies for each of the payoff differences between the options (the omitted category is 10 cents higher payoff for betting on tails)—and finds similar results.[15]

---

[14] While we urge caution in interpreting the finding due to the large amount of noise in participants' responses, the possibility that the gambler's fallacy is stronger for the 2nd-closest flip than the closest flip may be consistent with the conditional-probability evidence discussed above, which finds that participants think a tail is more likely than a head after HH but not after H. These findings are contrary to the formal gambler's-fallacy models in Rabin (2002) and Rabin and Vayanos (2010).

[15] Interpreting the coefficients on the controls, there is a nearly-significant 7-percentage-point positive coefficient on the dummy for whether the sequence was target-later, i.e., people on average think a head is more likely in the target-later questions, regardless of the outcomes of the other flips. We did not expect to find such an effect, and we have no explanation for it. The 2.5-percentage-point positive effect of the dummy for whether the betting-on-heads option appeared as the top option suggests that there is a small bias in favor of betting on the top option, but it is not statistically significant. The coefficients are essentially zero when betting on tails pays off 6 cents or 2 cents more than betting on heads, indicating that participants' behavior is not sensitive to how much more the tails option pays off. Relative to when the tails option pays off more,

To investigate forward-looking vs. backward-looking GF, columns 3 and 4 restrict the sample to target-later and target-earlier sequence pairs, respectively. Estimates are noisier because the sample is smaller. The 2nd-closest-flip effect appears to be stronger for target-later data, but we cannot reject the hypothesis that the coefficient on the 2nd closest flip is the same in column 4 as in column 3. Focusing on column 4, we find essentially no evidence supporting the existence of backward-looking GF. Overall, we interpret our evidence as suggestive that forward-looking GF is stronger than backward-looking GF. This asymmetry contradicts the formal models of the Rabin (2002) and Rabin and Vayanos (2010) models. 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).

Turning back to the streak data in Figure 5b, we assess the magnitude of GF by estimating the parameters α and δ of Rabin and Vayanos' (2010) model[16]:

$$q_t = \omega - \alpha \sum_{k=0}^{\infty} \delta^{k+1} y_{t-1-k} \quad (1)$$

where q_t represents the agent's belief regarding the t^{th} flip; $y_{t\text{-}1\text{-}k}$ represents the outcome of the $(t\text{-}1\text{-}k)^{th}$ 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

---

participants are 17-20% more likely to bet on heads when the heads option pays off more, but again, the likelihood of betting on heads is similar regardless of whether the difference is 2, 6, or 10 cents.

[16] Relative to equation (4) in Rabin and Vayanos (2010, p.736), our equation (1) is different in three ways. First, the variables that Rabin and Vayanos denoted ε_t and $\varepsilon_{t\text{-}1\text{-}k}$, we denote as q_t and $y_{t\text{-}1\text{-}k}$ because in our context we think it is clearer to distinguish notationally between an agent's perceived probability of a head (q_t) and the past outcome of a coin flip ($y_{t\text{-}1\text{-}k}$). Second, Rabin and Vayanos allow the bias of the coin to be a function of t, but we impose that is ω constant. Finally, we correct a typo by writing δ^{k+1} (rather than δ^k).

subsequent flips. The outcome variable $y_{t\text{-}1\text{-}k}$ is equal to +1 if the $(t\text{-}1\text{-}k)^{\text{th}}$ 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, $p_t = \frac{q_t+1}{2}$, is a rescaling of the belief variable $q_t \in [-1,1]$.

We conduct the estimation in Stata 10.1 using non-linear least squares on the first-difference of equation (1): $q_t - q_{t+1} = \alpha\delta^t$ (where every $y_{t\text{-}1\text{-}k}$ is equal to 1 because each realization of an outcome was a head in the sequences we presented to participants). The data we use are the mean judgments p_t shown in Figure 5b, as transformed by $q_t = 2p_t - 1$. The parameter estimates are $\hat{\alpha}$ = 0.031 (with a standard error of 0.022) and $\hat{\delta}$ = 0.947 (with a standard error of 0.152).[17] The positive point estimate for α is consistent with GF, but the 95% confidence interval includes zero. The point estimate for δ is close to one, indicating that there is no evidence for a decay of the GF in our data. This absence of decay is evident in Figure 5b as a decline in participants' probability that is nearly linear in m.

Next, we 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 in the betting task. For each bet made by each participant, we coded it as a 1 if it coincides with what a Rabin-Vayanos agent with α = 0.031 and δ = 0.947 would have done, and 0 otherwise. We estimated the amount of noise in participants' betting responses by calculating the variance σ^2 of the resulting binary variable. Then, we simulated 1,000 datasets by taking what the Rabin-Vayanos agent would have done when faced with the participants' options and adding i.i.d. mean-zero noise to the responses, drawn from a normal distribution with variance σ^2. Finally, we estimated the same regression as in column 1 of Table 3 but using each of the simulated datasets.

Column 5 of Table 3 reports the mean coefficients, averaged across the 1,000 simulations. 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. Keeping in mind that column 5 is an average across many realizations while column 1 is the result of a single realization (the actual sample of data we observed), the

[17] When we estimate this regression on the individual-level data, imposing the same parameter values for all individuals but with standard errors clustered by individual, the results are similar: $\hat{\alpha}$ = 0.031 (with a standard error of 0.022) and $\hat{\delta}$ = 0.947 (with a standard error of 0.152).

regression output is broadly similar. In column 5, the regression coefficients of each of the previous five flips is -0.045—the same order of magnitude as the coefficients in column 1—but the noise is sufficiently large that none of these coefficients is statistically significantly distinguishable from zero.[18]

To assess what the degree of GF we observe in the streak data would imply for our histogram elicitations, in each of Figures 2-5 and 7, we show with green lines what the Rabin-Vayanos agent would believe. The figures suggest that the magnitude of GF in the streak data would generate rather little deviation of beliefs from the true probabilities in the histogram elicitations.

For our first test of exact representativeness, we note that in Figure 3, the Rabin-Vayanos model alone fits the data quite well. Therefore, contrary to what exact representativeness would predict, there is no evidence that participants put extra weight on exactly 5 heads beyond the extra weight on the most likely outcomes generated by LSN. The second test is whether participants estimate a higher probability for the event "481-519 heads" in the elicitation in Figure 2 than for the event "451-549 heads" in Figure 7. The mean probability assigned to the former is 35.6%, while for the latter it is *higher* ($p < .01$) at 39.6%, contrary to the prediction.

Turning to the investigation of bin effects, the results are strong and easier to see. In the five-bin elicitation (shown in Figure 3), the mean probability assigned to 0-3 heads is 18%. As predicted by support theory, this is smaller than the sum of the probabilities in the eleven-bin elicitation (shown in Table 1) assigned to the constituent outcomes, 0, 1, 2, and 3 heads, which is 29.5% ($p < .01$). Similarly, participants assigned 14% to the event 7-10 heads in the five-bin elicitation, smaller than the sum in the eleven-bin elicitation, which is 25.5% ($p < .01$).[19] In the three-bin elicitation (Figure 4a), the average probabilities assigned to 0-4 and 6-10 heads are

---

[18] Naturally, since none of the other variables we manipulated—target earlier vs target later, the prize difference, and whether the heads option is listed on top—matters in the Rabin-Vayanos model, we do not include any of them as controls.

[19] In all of the histogram-elicitation questions, there is a tendency for participants to assign greater weight to the categories presented earlier (e.g., reporting greater probability for 0-3 than 7-10 heads). 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, 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.

34% and 30%, each significantly smaller than the sum of the average probabilities assigned to the individual outcomes, 42% and 39% ($p < .01$). When we asked participants to estimate *separately* the frequency of each of the eleven possible outcomes (Figure 6), each is judged more likely than in the 11-bin treatment and than the true probability (all at $p < .01$). For the sample size of 1,000, in the three-bin elicitation shown in Figure 7, participants estimated the probabilities of 0-450 and 550-1000 heads to be 32% and 28%, respectively. In the eleven-event elicitation described in Table 1c, the implied beliefs for these events are 46% and 36% (both significantly different at $p < .01$).

Recall that, as a training task, we asked participants to guess the percentage of the population in the United States composed of each of six major racial groups. While we did not formulate any research questions regarding these data ex ante, participants' responses seem to reflect bin effects. Figure 8 displays participants' mean estimates, 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 38.5%. The least frequent group is Native Americans, which comprise 0.7% of the population according to the Census but 7.5% according to participants.

Keeping in mind the strong evidence of bin effects in our data, we turn to evaluating the evidence about whether people exaggerate the likelihood of extreme proportions as predicted by NBLLN. Table 1 compares Kahneman and Tversky's (1972) evidence for sample sizes of 10 and 1,000 with ours. In some details, our findings differ. In particular, our participants put more probability mass in the bins presented to them earlier—for example, they assign higher probability mass to 0-5% heads than to 95-100% heads despite the symmetry—and participants put higher probability (4.2%) on the 95-100% bin in the sample of 1,000 than in the sample of 10 (2.8%; $p = .01$), and lower probability (6.1%) on the 0-5% bin in the sample of 1,000 than in the sample of 10 (9.2%; $p = .09$). 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 10 times or 1,000 times, Hence, overall, our findings qualitatively replicate Kahneman & Tversky's.

Taken at face value, these data would seem to provide strong evidence in favor of NBLLN. Both Kahneman & Tversky's data and our Table 1, however, confound the NBLLN overweighting of extreme outcomes with bin effects that bias all bins toward equal weight. For

sample size 1,000, however, we believe NBLLN is demonstrated. People think there is a 60.4% chance that the number of heads will be outside the range 451-549, even though the true probability is 0.2%. This seems unlikely to be due to compression. Could it be that subjects would accurately believe the true probability of 0 to 450 heads (say) is .1%, but due to bin effects report around 30%? While our evidence suggests that bin effects are powerful, we do not think they are this extreme. To see why that degree of compression seems inconsistent with our data, consider again the results shown in Figure 6. In that case where we asked subjects to assess frequency of the bins 0, 1, 2, 8, 9, and 10 in isolation, whose actual probabilities ranged from 0.1% to 4.4%, subjects reported between 18% and 28%. Note that each of these cases involve not only asking a category where all other possibilities are binned together, but also where the elicitation mode made the focus on the asked-about number of heads most salient—not even mentioning the other possibilities. Given that none of these categories induced stated frequencies above 28%, and the 0.1% categories only got bumped up to 18%, it is inconsistent that the 0.1% bins would induce reported beliefs averaging 30% in the 3-bin case due solely to bin-induced compression effects.[20]

Also inconsistent with NBLLN for $N = 1000$, Figure 2 shows that the mean probabilities assigned to the three outcomes, 0-480 heads, 481-519 heads, and 520-1000 heads, are close to equal: 35%, 36%, and 29% (these first two are statistically indistinguishable, but the first and second each differ from the third at $p = .06$ and $p = .04$, respectively). The corresponding true probabilities are: 11%, 78%, and 11% (all p's $< .01$). Compressing beliefs of 78% and 11% to near equality due solely to bin effects likewise seems implausible—although the isolated-category bins of 3 heads and 7 heads, each with true probability 11, in fact induce reports about 30%, close to the compression we see here. However, we suspect that that compression overstates the strength of bin effects we should observe in the 3-bin case for $N = 1000$ because it is with 2 bins (rather than 3) and with salient elicitation.

In contrast, for samples of size 10, we interpret our evidence as indicating that, once bin effects are controlled for, people's beliefs about the distribution of heads are reasonably well-

[20] Of course, the true probabilities in each bin are not the appropriate comparison in principle: we need to ask how subjects might be distorting their 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.

calibrated and in fact slightly overweight the likelihood of 5 heads. The data in Figure 1 and Table 1 would seem to suggest that, consistent with NBLLN, people substantially overweight the likelihood of extreme outcomes: e.g., the total reported probabilities of 8 or more heads out of 10 flips is 16.5%, compared with a true probability of 5.5%. But this apparent overweighting of extreme outcomes is due entirely to bin effects. The most telling evidence comes from the five-bin elicitation: 0-3, 4, 5, 6, and 7-10 heads, where the true probabilities of these outcomes are, respectively, 17.2%, 20.5%, 25%, 21%, and 17.2%. Since the true probabilities are roughly equal, bin effects per se are unlikely to have a large influence on participants' beliefs. Figure 3 shows participants' mean beliefs are 18.3%, 21.5%, 28.1%, 18.3%, and 13.8%, quite close to the correct probabilities (respective *p*'s are .45, .40, .03, .06, and .0006).

Recall from above that we find GF in the streak data in samples of size 10, which would lead to overweighting the likelihood of exactly 5 heads. We find consistent evidence in the three-bin elicitation, 0-4, 5, and 6-10 heads (seen in Figure 4a), with correct probabilities 38%, 25%, and 38%. Here, bin effects would predict that participants will underweight 5 heads by less than they should. Instead, we find that participants overweight 5 heads: mean probabilities are 34%, 36%, and 30%. Taken together, our results suggest that, for samples of size 10, NBLLN is actually somewhat outweighed by GF.

Our integrated design—with subjects responding to different questions about the very same set of flips—allows us to test in a much stronger way than is usually possible whether a single, quasi-Bayesian model could explain participants' reported beliefs. Of course, explaining the bin effects we document with a quasi-Bayesian model would be extremely difficult. Furthermore, even aside from bin effects, our results suggest that participants' beliefs about the likelihood of the total number of heads in the sample are inconsistent with their beliefs about specific sequences of flips. In the streak questions (Figure 5b), when asked how frequently 9 heads is followed by a head, participants yielded beliefs (per GF) that HHHHHHHHHH is half as likely as HHHHHHHHHT. And 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' sample 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 (Table 1), they reported that 9 heads is only 2.3 times more likely than 10 heads. The responses from Table 1 are influenced by the binning, of

course, so given our evidence that participants' sample beliefs are roughly calibrated correctly for a sample size of 10, we might instead consider the true ratio of the probabilities of 9 to 10, which is 10. Even this true ratio, however, is far smaller than the ratio of at least 20 implied by the sequence beliefs.

4. Discussion and Conclusion

We draw three main conclusions from our experiment. First, we find evidence for three major biases in beliefs—GF, NBLLN, and bin effects—in an incentivized experiment. The GF we find is weaker than what we anticipated based on results from prior research, while the bin effects are stronger. Second, we find that bin effects are an important potential confound that need to be accounted for in designing experiments about and interpreting evidence on sample beliefs. In particular, while previous work has concluded that beliefs about the proportion of heads in a sample are the same across samples of size 10 and 1,000, we find that the results for a sample size of 10 are instead driven by bin effects. Aside from bin effects, our participants exhibit reasonably well-calibrated beliefs, somewhat overweighting the most likely outcomes as per GF. Third, our evidence suggests that participants' sample beliefs and sequence beliefs cannot be reconciled by a single, incorrect model of the data-generating process to which the rules of probability are applied correctly.

Our results have several implications for theoretical work aimed at modeling belief biases. Theoretical efforts to model bin effects (notably Tversky and Koehler (1994) and Ahn and Ergin (2010)) have assumed that these effects apply to *subjective* probabilities, and do not emphasize the same possibility for objective probabilities. Our evidence comes from coin flips known to be fair, however, and hence existing models fall short of capturing the range of circumstances to which bin effects apply. Most economic models to date of belief biases have attempted to reconcile different biases within a quasi-Bayesian framework. Our finding that participants' sample beliefs and sequence beliefs are inconsistent suggests that quasi-Bayesian modeling approaches will not be able to predict some important aspects of people's beliefs.

Despite our own surprise at how strong they are, our findings that bin effects are powerful echo results from previous work such as Tversky and Koehler (1994) and Fox and Clemen

(2005). Nonetheless, we are not aware of any work in economics or psychology focused on measuring people's beliefs that accounts for bin effects interpreting the data.[21] However, we suspect that bin effects likely matter a great deal for drawing appropriate conclusions in a wide range of contexts. For example, 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), and about distribution of equity returns (e.g., Dominitz and Manski, 2007). These literatures tend to find that people's belief distribution biased in the direction of uniform, but bin effects alone could generate that pattern.

In addition to the specific insights we obtained about the sample beliefs, we think the integrated feature of our experimental design prove useful more generally as a methodological tool for investigating biases that seem contradictory. While sequence and sample beliefs are usually investigated in separate studies, by examining them jointly with respect to beliefs over the very same set of sample realizations, we were able to investigate whether these beliefs are mutually compatible.

We have attempted to be fully transparent in communicating our ex ante research questions that motivated our experimental design. Having run the experiment—and having seen some surprising results—it is now clear that there are many ways in which it falls short and which future research should improve upon. Given the noisiness of the betting data, it would have been better to ask less complicated questions, with all but one of the flips known. When eliciting conditional probabilities, rather than asking only about the likelihood of a head following a streak, if we had asked about other preceding sequences, we could have generated more situations in which we could examine direct inconsistencies across sequence and sample beliefs. Although we believe that bin effects alone cannot plausibly explain the overweighting of extreme outcomes that we observe for a sample size of 1,000, we could test this better both with better design (such as binning the distribution so that participants put equal weight on all the bins) and framed with ranges that provide higher-powered tests given the levels of NBLLN and bin effects. Finally, although we tried to mitigate the possibility of inference by subjects

[21] Indeed, the same brilliant researcher who documented striking evidence both of NBLLN and of bin effects, Amos Tversky, does not seem to have recognized that bin effects confound the interpretation of the data on NBLLN.

confounding the interpretation of results, the software we used permitted randomization of question order but not the ability to record that order. If we observed the order we could check whether it appears that subjects are initially inferring the process in a way that confounds our interpretations. Seeing whether range choices would allow us to identify obvious order effects.

We plan to run a further experiment to help fix some of these problems and to provide further insights on the topics explored in this paper. We intend to post our experimental design, research questions, and intended analysis on our websites prior to running it.

References

Barberis, Nicholas, Andrei Schleifer, and Robert Vishny (1998). "A model of investor sentiment." *Journal of Financial Economics*, 49(3), 307-343.

Benjamin, Daniel J., Collin Raymond, and Matthew Rabin (2011). "A Model of Non-Belief in the Law of Large Numbers." Cornell University mimeo, October.

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.

Griffin, Dale, and Amos Tversky (1992). "The Weighing of Evidence and the Determinants of Confidence." *Cognitive Psychology*, 24(3), 411-435.

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.

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.

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.

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

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.

Winkler, Robert L., and Allan H. Murphy (1973). "Experiments in the Laboratory and the Real World," *Organizational Behavior and Human Performance*, 10, 252-270.

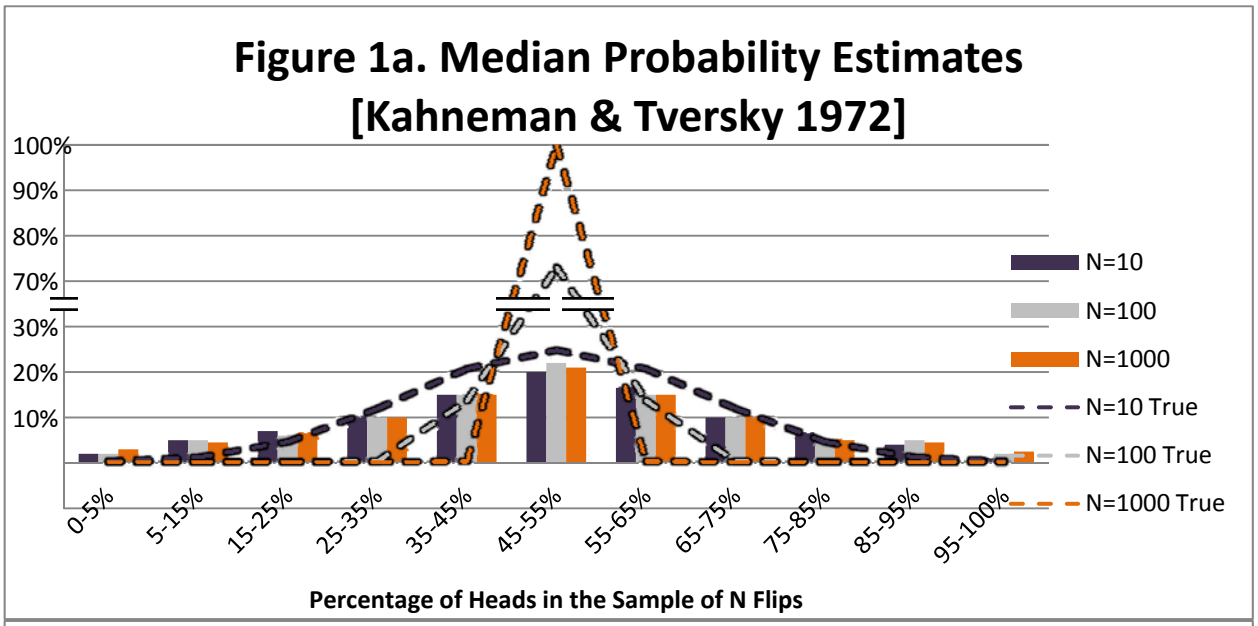


Figure 1a. Median Probability Estimates
[Kahneman & Tversky 1972]
100%
90%
80%
70%
30%
20%
10%
0-5%
5-15%
15-25%
25-35%
35-45%
45-55%
55-65%
65-75%
75-85%
85-95%
95-100%
N=10
N=100
N=1000
N=10 True
N=100 True
N=1000 True
Percentage of Heads in the Sample of N Flips

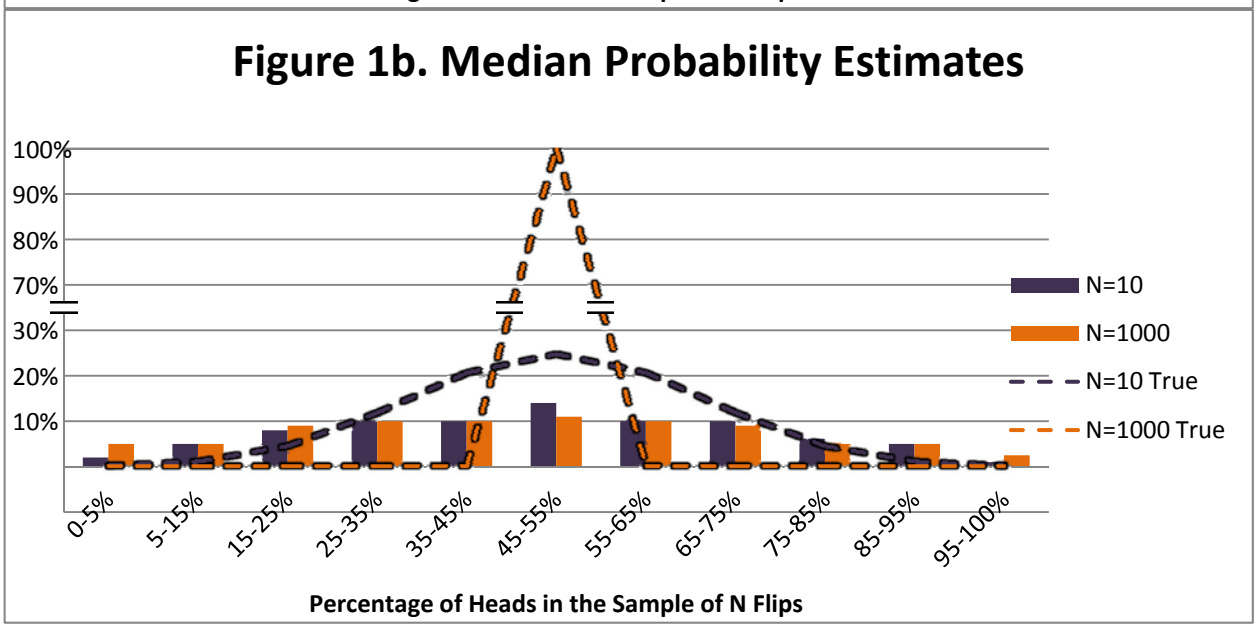


Figure 1b. Median Probability Estimates
100%
90%
80%
70%
30%
20%
10%
0-5%
5-15%
15-25%
25-35%
35-45%
45-55%
55-65%
65-75%
75-85%
85-95%
95-100%
N=10
N=1000
N=10 True
N=1000 True
Percentage of Heads in the Sample of N Flips

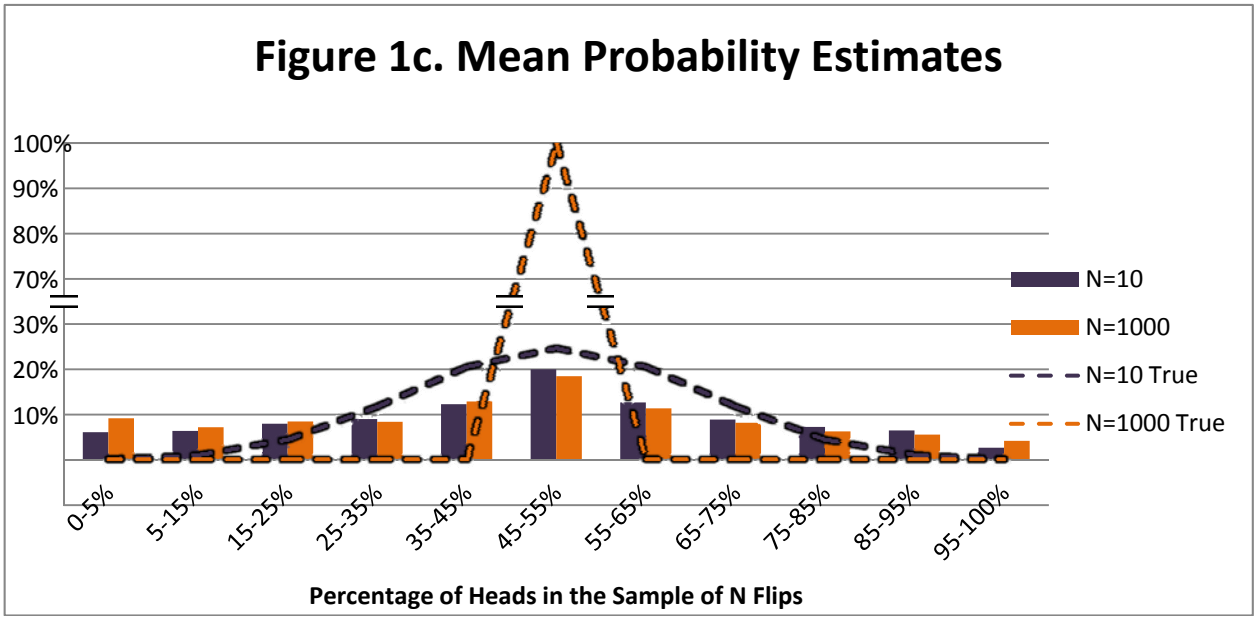


Figure 1c. Mean Probability Estimates
100%
90%
80%
70%
30%
20%
10%
0-5%
5-15%
15-25%
25-35%
35-45%
45-55%
55-65%
65-75%
75-85%
85-95%
95-100%
N=10
N=1000
N=10 True
N=1000 True
Percentage of Heads in the Sample of N Flips

Figure 2

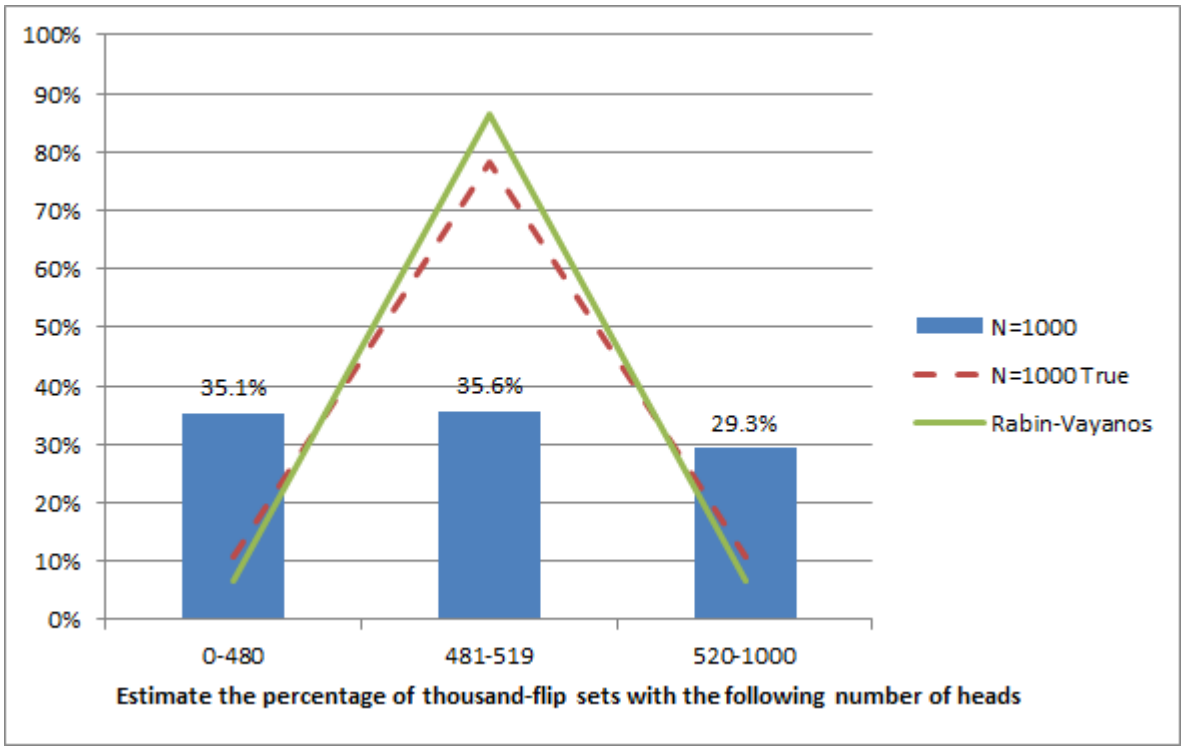

100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
35.1%
35.6%
29.3%
0-480
481-519
520-1000
N=1000
N=1000 True
Rabin-Vayanos
Estimate the percentage of thousand-flip sets with the following number of heads

Figure 3

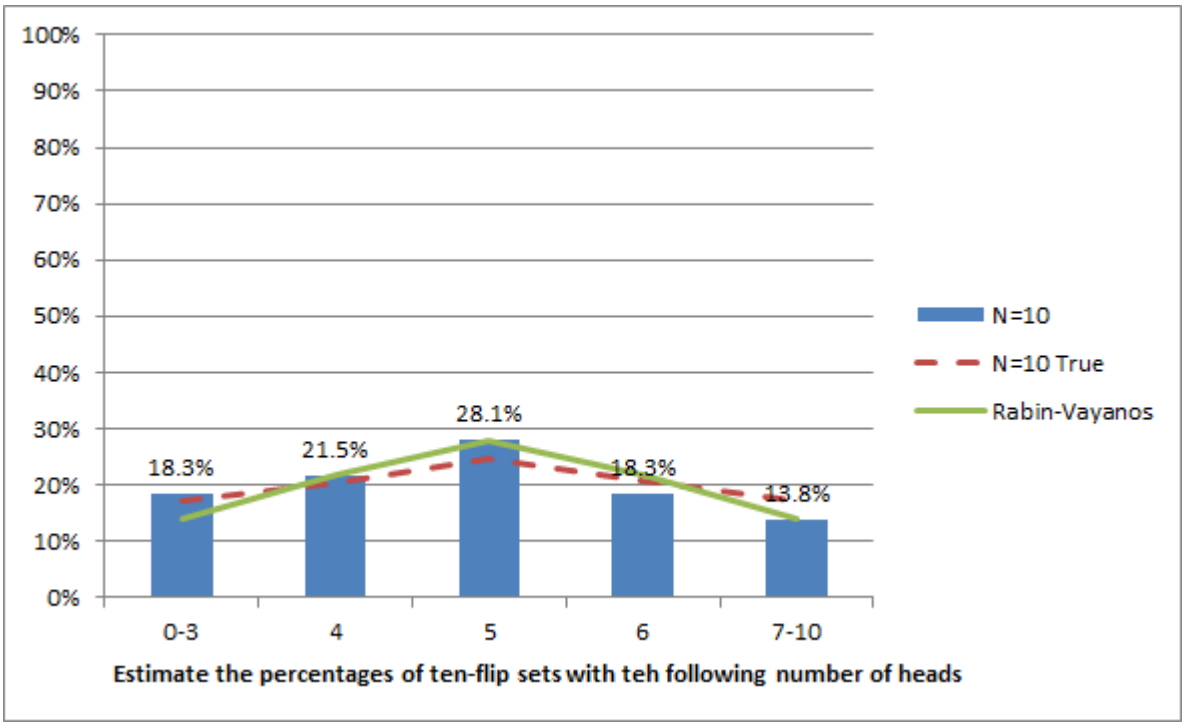

100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
18.3%
21.5%
28.1%
18.3%
13.8%
0-3
4
5
6
7-10
N=10
N=10 True
Rabin-Vayanos
Estimate the percentages of ten-flip sets with teh following number of heads

Figure 4a

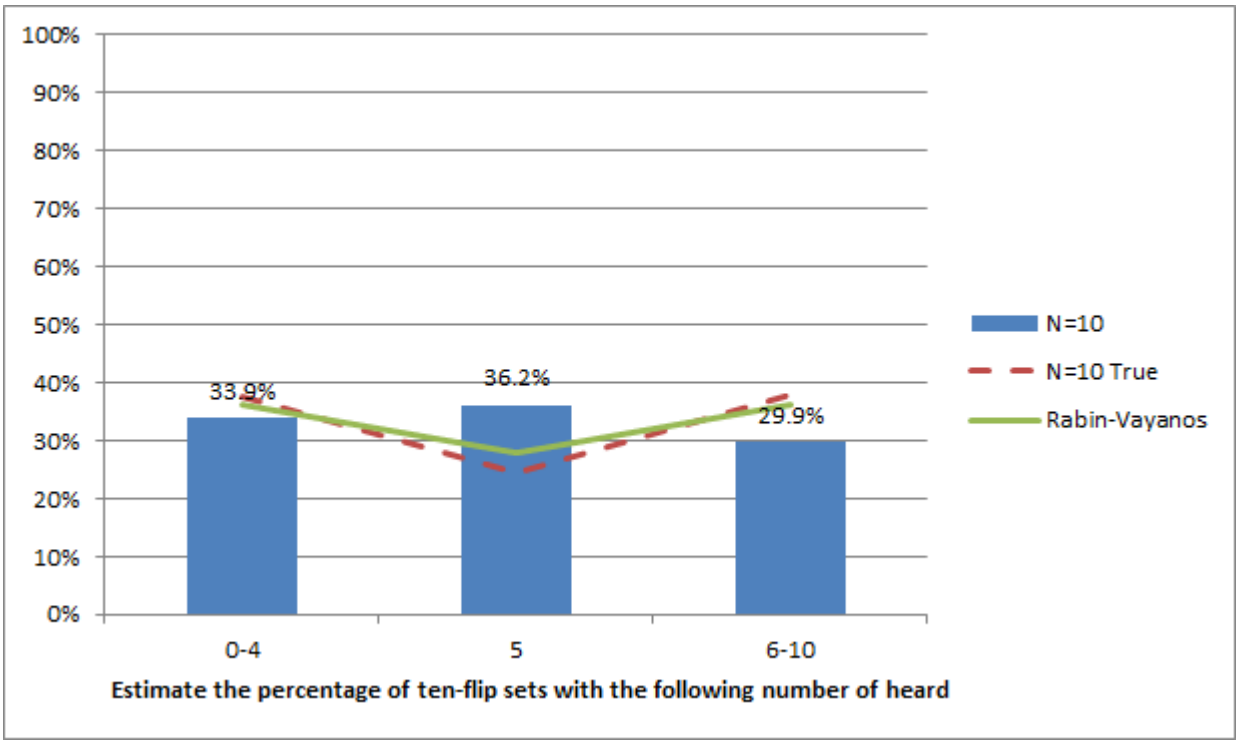


Figure 4b

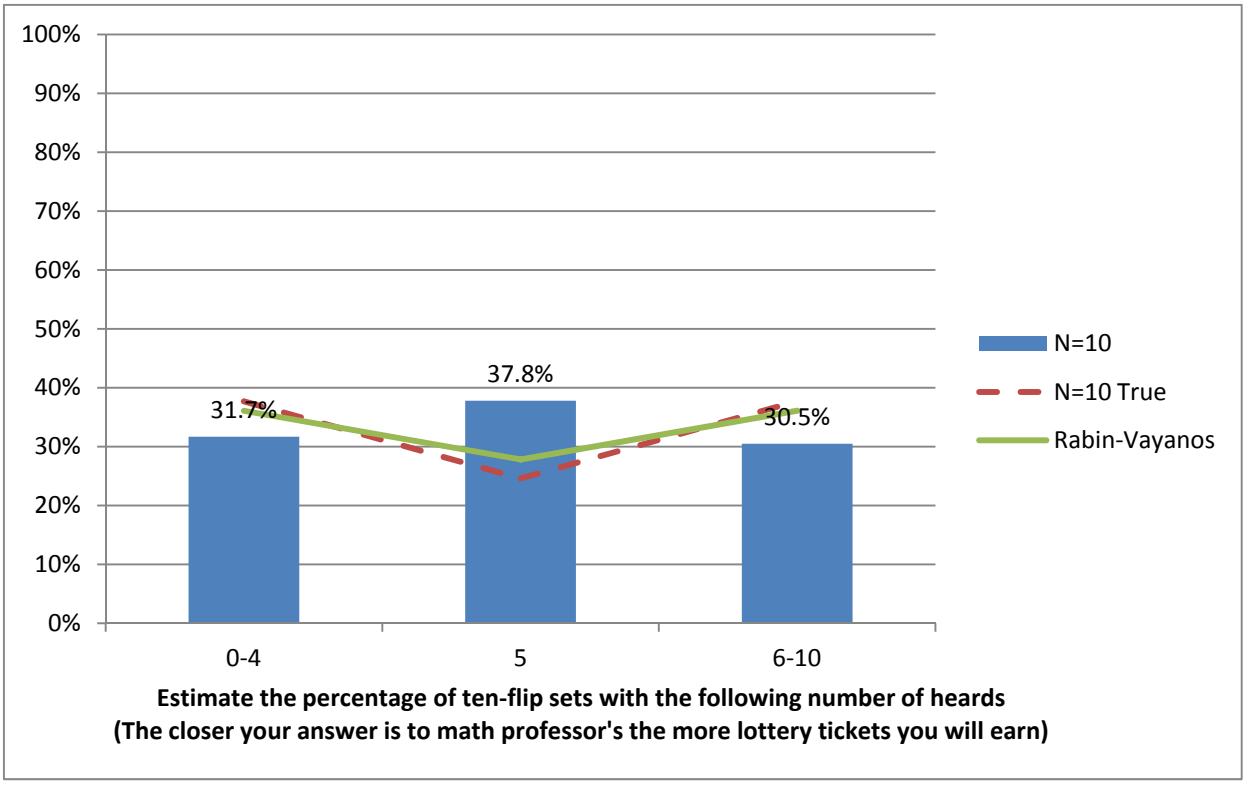


Figure 5a

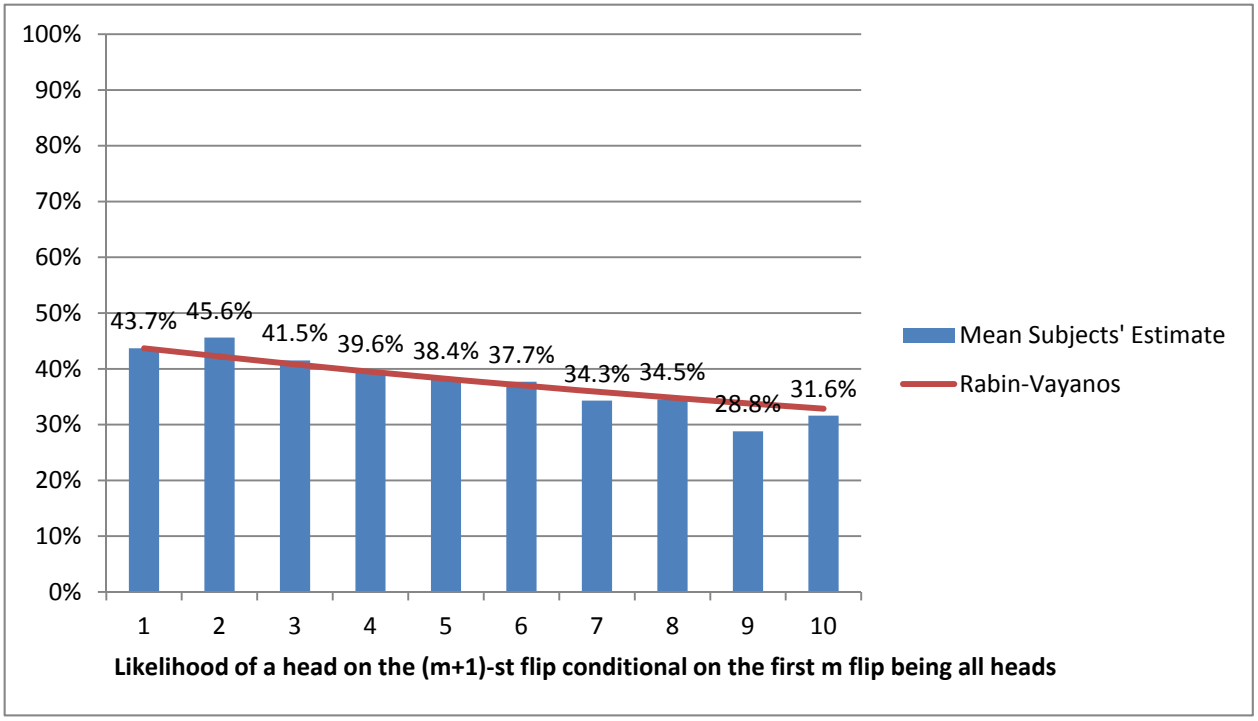


Figure 5b (dropping confused subjects)

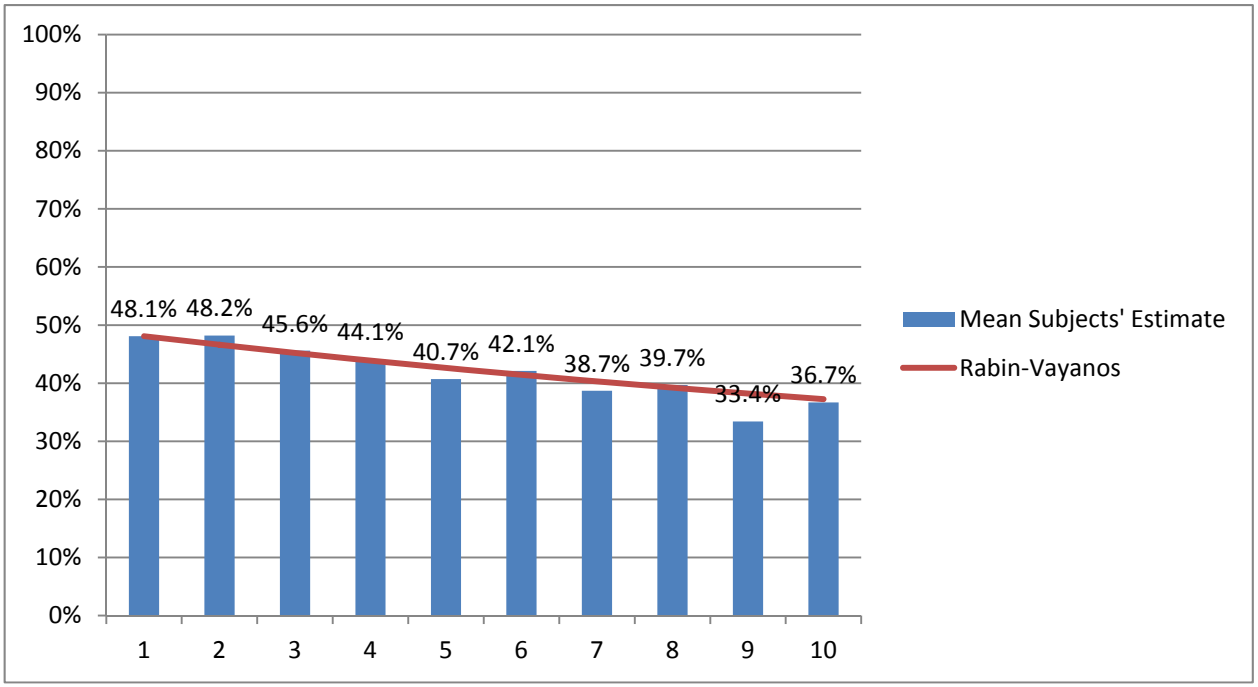


Figure 6

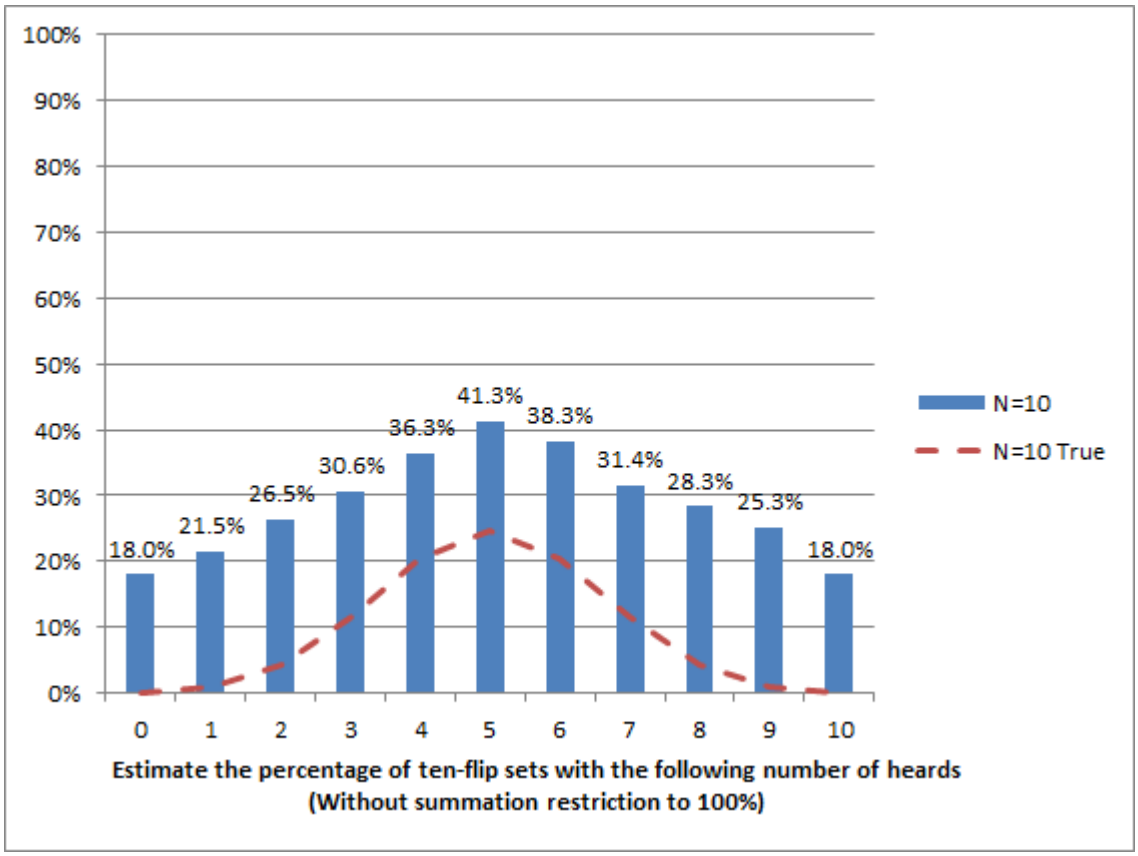

100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
18.0%
21.5%
26.5%
30.6%
36.3%
41.3%
38.3%
31.4%
28.3%
25.3%
18.0%
0
1
2
3
4
5
6
7
8
9
10
N=10
N=10 True
Estimate the percentage of ten-flip sets with the following number of heards
(Without summation restriction to 100%)

Figure 7

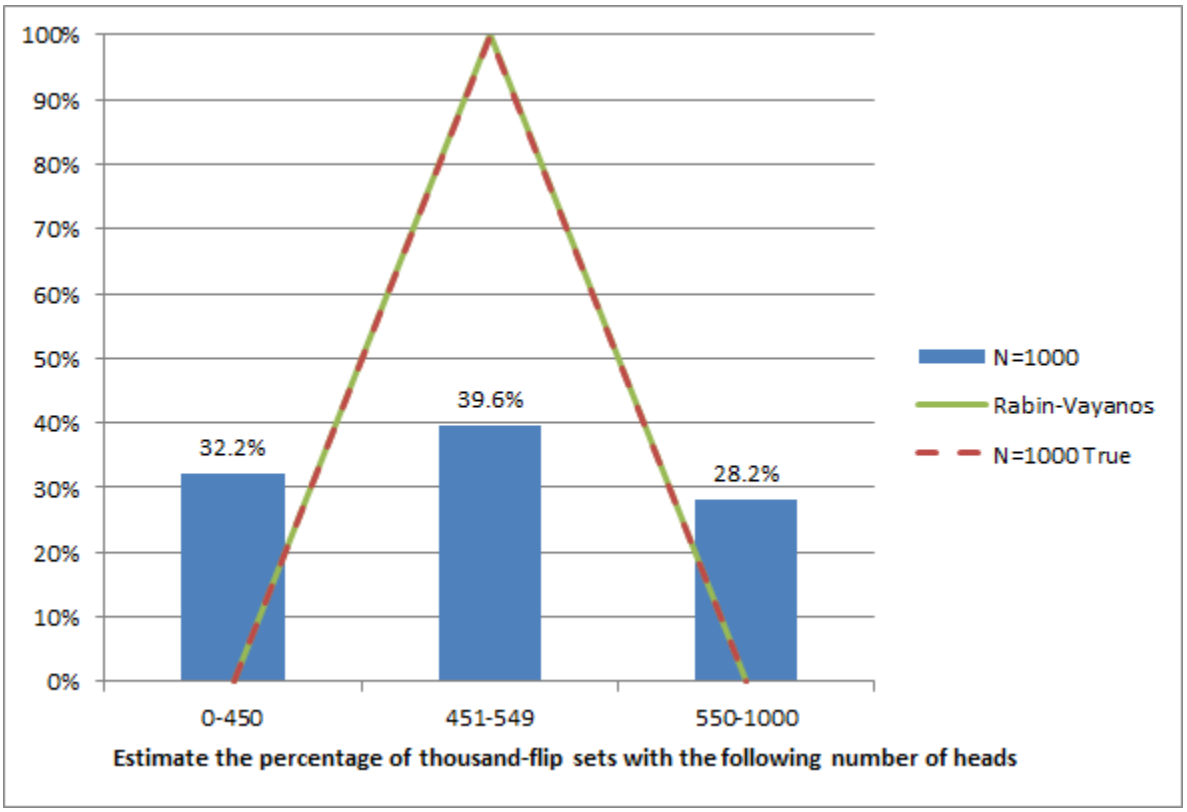

100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
32.2%
39.6%
28.2%
0-450
451-549
550-1000
N=1000
Rabin-Vayanos
N=1000 True
Estimate the percentage of thousand-flip sets with the following number of heads

Figure 8

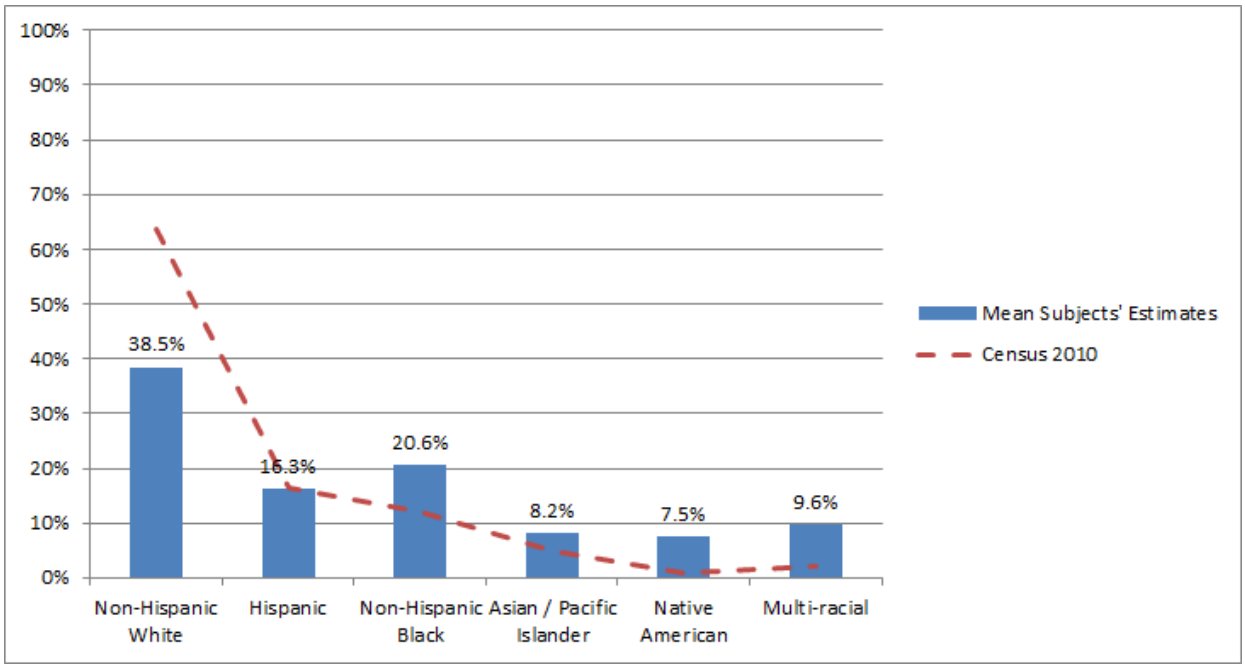

100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
38.5%
16.3%
20.6%
8.2%
7.5%
9.6%
Non-Hispanic White
Hispanic
Non-Hispanic Black
Asian / Pacific Islander
Native American
Multi-racial
Mean Subjects' Estimates
Census 2010

Table 1: Histogram beliefs for samples of 10 and 1000

Sample size of $N = 10$					Sample size of $N = 1000$				
Heads	Correct	K&T: Medians	Our Data: Medians	Our Data: Means	Heads	Correct	K&T: Medians	Our Data: Medians	Our Data: Means
0	0.1%	2.0%	2.0%	6.1%	≤5%	0.0%	3.0%	5.0%	9.2%
1	1.0	5.0	5.0	6.4	5-15	0.0	5.0	5.0	7.1
2	4.4	7.0	8.0	8.0	15-25	0.0	7.0	8.5	8.5
3	11.7	10.0	10.0	9.0	25-35	0.0	10.0	10.0	8.4
4	20.5	15.0	10.0	12.3	35-45	0.1	15.0	10.0	13.0
5	24.6	20.0	14.0	20.0	45-55	99.8	21.0	11.5	18.7
6	20.5	17.0	10.0	12.7	55-65	0.1	15.0	10.0	11.4
7	11.7	10.0	10.0	8.9	65-75	0.0	10.0	9.0	8.2
8	4.4	7.0	6.0	7.3	75-85	0.0	5.0	5.0	6.3
9	1.0	4.0	5.0	6.5	85-95	0.0	5.0	5.0	5.1
10	0.1%	1.0%	1.0%	2.8%	≥95%	0.0%	3.0%	2.0%	4.2%

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

Table 2: Descriptive statistics on participants' bets

Number of heads in Sequence	Full Sample	Target-later	Target-earlier
0	60.9% [23]	66.7% [12]	54.6% [11]
1	59.8% [164]	68.0% [78]	52.3% [86]
2	47.6% [313]	47.8% [148]	47.3% [165]
3	49.7% [306]	53.1% [164]	45.8% [142]
4	47.3% [146]	47.9% [71]	46.7% [75]
5	60.0% [35]	53.9% [13]	63.6% [22]
Total:	51.0% [987]	53.5% [486]	48.5% [501]

Note: Percentage of participants who chose H after a given number of H observed in sequence. The number of participants from whom the percentage is calculated is displayed in brackets.

Table 3: Regression results for participants' bets and simulated behavior from the Rabin-Vayanos model

	(1) Full Sample	(2) Full Sample	(3) Target-later	(4) Target-earlier	(5) RV Simulation
Closest Flip is H	-0.0095	-0.0095	-0.0497	0.0234	-0.0450
	(0.0326)	(0.0329)	(0.0493)	(0.0424)	(0.0382)
2nd Closest Flip is H	-0.0755*	-0.0720*	-0.1125**	-0.0398	-0.0452
	(0.0319)	(0.0320)	(0.0428)	(0.0453)	(0.0381)
3rd Closest Flip is H	-0.0196	-0.0181	-0.0134	-0.0185	-0.0443
	(0.0360)	(0.0351)	(0.0471)	(0.0501)	(0.0382)
4th Closest Flip is H	-0.0460	-0.0434	-0.0842	-0.0128	-0.0445
	(0.0329)	(0.0316)	(0.0505)	(0.0460)	(0.0382)
5th Closest Flip is H	0.0435	0.0411	0.0418	0.0365	-0.0470
	(0.0293)	(0.0297)	(0.0439)	(0.0392)	(0.0382)
Target-later		0.0679			
		(0.0351)			
Head Option on Top		0.0249	0.0174	0.0324	
		(0.0289)	(0.0484)	(0.0469)	
Prize Difference = -6		-0.0009	-0.0340	0.0399	
		(0.0535)	(0.0799)	(0.0769)	
Prize Difference = -2		-0.0090	-0.0237	0.0176	
		(0.0533)	(0.0732)	(0.0853)	
Prize Difference = 2		0.1820**	0.1470	0.2254**	
		(0.0665)	(0.0918)	(0.0820)	
Prize Difference = 6		0.1678**	0.1580	0.1875*	
		(0.0629)	(0.0823)	(0.0876)	
Prize Difference = 10		0.2020**	0.1288	0.2659**	
		(0.0640)	(0.0915)	(0.0917)	
Constant	0.5624**	0.4207**	0.5771**	0.3390**	0.6127**
	(0.0383)	(0.0620)	(0.0882)	(0.0792)	(0.0472)
Number of choices	987	987	486	501	682

Notes: Linear probability regressions. The dependent variable in columns 1-4 is a dummy for betting on the heads sequence. The independent variables are dummies for each of the five flips closest to the target flip being a head, a dummy for whether the sequence was target-later, a dummy for whether the betting-on-heads option appeared on the top of the screen, and dummies for each of the payoff differences between the betting-on-heads option and the betting-on-tails option (the omitted category is -10 cents). Standard errors are clustered by participant. Columns 1 and 2 show results for the full sample. Columns 3 and 4 restrict the sample to target-later and target-earlier sequence pairs, resp. As described in the text, column 5 displays average results from 1000 simulations, with the dependent variable generated as what the participants would have chosen if they behaved according to the Rabin-Vayanos model (with $\alpha = 0.031$, $\delta = 0.947$), observed the same 5 coin flips as actual participants did, and had the same residual variance as the actual participants. These standard errors are adjusted for variance in the coefficients across simulations. * $p < 0.05$, ** $p < 0.01$

Appendix A: Ancillary Questions and Results

In this Appendix we present all the remaining research questions and hypotheses that we formulated prior to running the experiments, jointly with the relevant results. In light of those questions, we also explain additional aspects of the experimental design.[22] Although most of our research questions were clearly old or new "hypotheses"—questions for which we had ex ante strong beliefs about either what we believed or what would constitute answers consistent with previous theory—some of our concrete questions were more speculative, where we either had no strong ex ante beliefs or were looking for relatively unlikely alternative possible facets of participants' beliefs.

Whether bets are affected by payoff differences. Note that, in principle, participants' predictions about the likelihood of a head should be unrelated to the magnitude of the payoff difference between the options: since we are asking about which is more common among one million samples, a person should bet on anything they think is above 50% irrespective of payoffs and risk attitudes. Ironically, participants who do not believe in the law of large numbers might not do so, but we did not think this is a likely influence here and did not test it. In principle, some forms of sensitivity to payoffs may indicate NBLLN for sample sizes of one million; but we believe less interpretable noise (of the sort virtually always observed in experiments) is a more likely explanation for even instances that look that way. The reason is that, given their (possibly mistaken) views about random sequences of coin flips, a participant will have some prediction about whether heads or tails is more likely, and will prefer to bet on whichever option gives a higher expected value. In table A, we test whether predictions about the likelihood of a head are related to the payoff differences. To make comparison more convenient, column 1 is identical to Table 3's column 1. Table A's columns 2, 3, and 4 restrict the sample to small payoff differences (2 cents in favor of either heads or tails), medium payoff differences (6 cents), and large payoff differences (10 cents), respectively. Across the fifteen pairwise tests of equality of

[22] This includes cases where our design, after the fact, was clearly not optimal to address the questions at hand.

coefficients across these columns, only one is statistically significant at the 5% level (the 4th closest flip between columns 6 and 7) to which we attach little meaning. Focusing on the 2nd-closest-flip effect, the point estimate is largest for small payoff differences. In this case, participants are 13 percentage points less likely to bet on the heads option when the 2nd closest flip to the target was a head rather than a tail. The point estimate is smaller the larger the payoff differences between the options, though due to the large standard errors, we cannot draw strong conclusions.

Whether people are aware of the correct probabilities. The most likely interpretation of judgmental biases studied here, and more generally, is that people do not know that they are considered wrong. But we wondered—inspired by some anecdotes—whether it might be the case that some participants were aware of but disapproving of the (high-falutin) ideas about probabilistic judgments that "authorities" had. The beginning of our post-experimental questionnaire was designed to address this research question. We told participants:

> "For the following questions, please answer the way a mathematics professor would answer them. In fact, we asked a mathematics professor to answer all of these questions. We will reward you for answering the way the math professor answered. All of the questions pertain to the million ten-flip sets."[23]

We then asked participants how they thought the math professor answered these four questions:

(1) The histogram of the frequency of 0-4, 5, and 6-10 heads. The mean probabilities are displayed in Figure 4b, none of which differs significantly from participants' own beliefs shown in Figure 4a.

---

[23] In fact, we did ask a prominent math professor these questions, but we did not do so until after the experiment was completed. His answers were correct: (1) 37.7%, 24.6%, and 37.7%; (2) 50%; (3) the pattern that pays 51 tickets; and (4) 50%.

(2) The frequency of the third flip being a head when the first two were heads; and the frequency of the tenth flip being a head when the first nine were heads. The mean estimates were 40.9% and 37.1%, respectively, which are both significantly smaller than 50% (both p's < .01), and the latter estimate is marginally smaller than the former ($p < .09$). Hence participants apparently think that even math professors believe in LSN. Participants' estimates on these questions are similar to their estimates when they are reporting their own beliefs (rather than the math professor's)—41.5% and 31.6%, respectively—although the latter is significantly smaller than the corresponding estimate of 37.1% for the math professor ($p = .03$).

(3) Whether to bet that xxHHHHHxTx was more frequent, which pays off 49 tickets if correct, or that xxHHHHHxTx was more frequent, which pays off 51 tickets if correct. 68% of participants (significantly more than 50%; $p = .0002$) thought that the math professor was more likely to bet on the former, consistent with GF. In contrast, when facing this bet themselves, only 40% bet on the former sequence, as shown in Table 2. This difference might mean that participants think a math professor is more likely to believe in GF. An alternative interpretation that we find plausible is that the math professor version of this question is less noisy because participants were thinking about only 1 bet, rather than 10; under that interpretation, the professor data may reveal clearer evidence of GF than is apparent from the data on participants' own bets in Table 2.

Taken together, the evidence from the professor questions suggests that people are unaware that their beliefs differ from what would be considered correct probabilities.

Whether belief biases are related to mathematical ability. Next in the post-experimental questionnaire, we asked participants six questions (listed in Appendix B) designed to test their facility with numbers and basic arithmetic. Our research question was whether people with greater math ability are less likely to exhibit GF, NBLLN, and bin effects.

Above-median math scorers—the 46% of participants who got more than 3 or more correct out of 6 on the math quiz—exhibit less NBLLN. On the eleven-bin histogram for 1000-flip sets, while below-median math-scorers assign 14.3% probability to the event "45-55% heads," above-median math-scorers assign a significantly higher probability of 24.0% ($p < .01$),

although both groups exhibit significant NBLLN relative to the true probability of 99.8%. On the five-bin histogram for 10-flip sets, the same pattern is evident, although below-median math-scorers actually come closer to estimating the true probability of the event "5 heads" than above-median math-scorers. Below-median math-scorers' mean estimate is 25.1% (not significantly different from the true probability), while above-median math-scorers' mean estimate is 32.0%, significantly different from the below-median math-scorers' mean (p = .04) and from the true probability (p < .01).

There is little evidence of any relationship between math score and propensity to commit the gambler's fallacy. The coefficient from a regression of the estimated conditional probability of a head on the number of prior heads is -1.5 for the below-median math-scorers and -2.0 for the above-median math-scorers, but these are not at all statistically distinguishable (p = .47).

Above-median math-scorers may be less prone to bin effects. To assess this for 1000-flip sets, we subtract the reported probability for the event "45-55% heads" for the eleven-bin histogram from the reported probability for the same event for the three-bin histogram; the larger the discrepancy, the larger the bias caused by increasing the number of bins. While this difference is 24.4% for below-median math-scorers, it is a statistically significantly smaller 14.1% for above-median math-scorers (p = .03). To measure the bias for 10-flip sets, we analogously subtract the reported probability for the event "5 heads" for the eleven-bin histogram from the reported probability for the same event for the three-bin histogram. While the bias is again smaller for above-median math-scorers (14.2%, as compared with 17.8% for below-median math-scorers), this difference is not statistically significant. Furthermore, for below-median scorers, the sum of the probability estimates for 0, 1, 2, …, 10 heads (when these estimates are unconstrained) is 394%, and for high scorers, 282%. While both are much larger than 100%, this measure is marginally statistically significantly smaller for high scorers than for below-median scorers (p = .05).

In the 10-flip sets, there is a negative relationship between NBLLN and the gambler's fallacy for both above-median math-scorers and below-median math-scorers. The magnitude of this relationship, however, is greater for below-median math-scorers and only reaches statistical significance for below-median math-scorers.

Within-subject inconsistency. One of our hypotheses was that participants who exhibited greater gambler's fallacy would also exhibit greater NBLLN. A measure of a participant's degree of gambler's fallacy is (as mentioned in section 3) 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 participant's gambler's fallacy. As noted above, 57% of participants have a negative slope. A measure of a participant's degree of NBLLN in the 10-flip sets is the reported probability of "5 heads" in the five-bin histogram eliciting beliefs over the events 0-3, 4, 5, 6, and 7-10 heads. We choose this histogram question for our measure of NBLLN because about half (47%) of participants reported a probability less than the true value of 24.6%, indicating the presence of NBLLN. The correlation between these two variables is -0.24 (p = .02), suggesting that participants who exhibit a stronger gambler's fallacy tend to exhibit a weaker NBLLN, the opposite of our hypothesis. We also examine the correlation between a dummy variable for presence of the gambler's fallacy (slope less than zero) and a dummy variable for presence of NBLLN (reported probability of "5 heads" less than the true probability). This correlation is -0.11 (p = .26), suggesting that our conclusion is not driven by extreme values of our gambler's fallacy and NBLLN measures.

A measure of a participant's degree of NBLLN in the 1000-flip sets is the reported probability of "45-55% heads" in the eleven-bin histogram. The correlation between this measure of NBLLN and our regression-based measure of gambler's fallacy in the 10-flip sets is -.11 (p = .28). This relationship is weaker but in the same direction as when we use a measure of NBLLN in the 10-flip sets.

Heads bias and order effects. We checked whether presentation in terms of heads and the ordering of the questions—and we found nothing in particular.

To allow us to test whether there is a bias toward thinking heads is more likely than tails (and to neutralize such a bias by pooling the data across both frames), we randomized whether the question is framed in terms of heads or tails. For each participant for each histogram question, we also independently randomized whether the bins were labeled in terms of the number of "heads" out of *N* flips or the number of "tails," and whether moving from left to right the bins were increasing in the number of heads/tails or decreasing. As expected, a few of the

comparisons are statistically significant, but none of these manipulations systematically affects participants' responses.

To examine (and neutralize) order effects, half the participants randomly faced the sections of the experiment in one order—betting questions, 10-flip questions, then 1000-flip questions—and half in another order—1000-flip questions, betting questions, then 10-flip questions. We can examine whether fatigue affected reported beliefs by comparing responses for the 1000-flip questions among participants who faced those questions first with responses among participants who faced those questions last. In the eleven-bin histogram for the 1000-flip set, the difference in the mean probability assigned to "451-549 heads" between these two groups of participants is tiny and statistically insignificant.

For the histogram questions, we also randomized whether the screen asked participants to guess the frequency in increasing or decreasing order of heads, and we find no evidence of this ordering influencing behavior.

Whether belief biases are related to demographics. As mentioned above, we asked participants their gender, age, and annual income (specified with a drop-down menu and five income categories: $0-$50K; $50K to $100K; $100K to $150K; $150K to $200K; $200K or above) in the final part of the post-experimental questionnaire. We asked these questions to investigate whether there were differences across demographic groups in the extent to which they exhibited the belief biases.

There are few systematic individual differences in participants' behavior by sex, age, income, and self-reported effort. The few exceptions—none ex ante anticipated relationships—are as follows. When we compare participants older than age 27 (47% of the sample) with younger participants, the older participants exhibit weaker support-theory "binning" effects, less NBLLN, and more gambler's fallacy, although this last relationship is less robust statistically. The negative relationship between NBLLN and gambler's fallacy is stronger and statistically significant for participants who reported "3" or less on a 5-point scale of effort (37% of the sample) than for those who reported "4" or more.

Table A: Regression results for participants' bets depending on prize differences

	(1) Full Sample	(2) Prize Diff.=±2	(3) Prize Diff.=±6	(4) Prize Diff.=±10
Closest Flip is H	-0.0095	-0.0119	-0.0047	-0.0171
	(0.0326)	(0.0527)	(0.0552)	(0.0555)
2nd Closest Flip is H	-0.0755*	-0.1293*	-0.0553	-0.0407
	(0.0319)	(0.0546)	(0.0586)	(0.0639)
3rd Closest Flip is H	-0.0196	-0.0838	0.0000	0.0322
	(0.0360)	(0.0556)	(0.0544)	(0.0631)
4th Closest Flip is H	-0.0460	-0.0925	-0.1181*	0.0720
	(0.0329)	(0.0561)	(0.0548)	(0.0607)
5th Closest Flip is H	0.0435	-0.0690	0.0583	0.1391*
	(0.0293)	(0.0541)	(0.0568)	(0.0581)
Target-later		0.0480	0.0870	0.0337
		(0.0520)	(0.0538)	(0.0598)
Head Option on Top		-0.0144	0.0528	0.0586
		(0.0562)	(0.0470)	(0.0501)
Constant	0.5624**	0.6830**	0.4880**	0.3776**
	(0.0383)	(0.0699)	(0.0846)	(0.0806)
Number of choices	987	347	342	298

Notes: Linear probability regressions. The dependent variable is a dummy for betting on the heads sequence. The independent variables are dummies for each of the five flips closest to the target flip being a head, a dummy for whether the sequence was target-later, and a dummy for whether the betting-on-heads option appeared on the top of the screen. Standard errors are clustered by participant. Column 1 shows results for the full sample and is identical to table 3's column 1. Columns 2, 3, and 4 restrict the sample to small payoff differences (2 cents in favor of either heads or tails), medium payoff differences (6 cents), and large payoff differences (10 cents), resp. * $p < 0.05$, ** $p < 0.01$

Appendix B

Math quiz

1) If a = 7b, then what does (5a)/7 equal?
2) If 2/6 = 2/(3+x), hat is the value of x?
3) It takes a small plane an hour to travel 144 miles. How many miles does it travel in 5 minutes?
4) A dog eats ¾ of a pound of meat every day. How many pounds of meat does the dog eat per week?
5) If 8x + 2y = 46, and y=7, what is x?
6) The volume of a cube is 27 cubic inches. What would be the volume of a cube whose sides were twice as long?