

A Model of Non-Belief in the Law of Large Numbers

Daniel J. Benjamin
Cornell University

Matthew Rabin
University of California — Berkeley

Collin Raymond*
University of Michigan

October 16, 2011

Abstract

Psychological research suggests that people believe that even in very large samples, proportions might depart significantly from the population mean. We model this “non-belief in the Law of Large Numbers” by assuming that a person believes that proportions in any given sample of binary signals might be determined by a rate different than the true rate. In inference, a non-believer attends too little to sample size, and remains uncertain even after observing an arbitrarily large sample. Non-belief is often a necessary enabler of other biases, such as the over-influence of vivid signals, that would otherwise be overwhelmed by the logic of the Law of Large Numbers. Because it prevents recognition that a large number of independent better-than-fair bets is extremely unlikely to yield an aggregate loss, non-belief helps explain how loss aversion can induce reluctance to accept even negligible risk. We explore various assumptions about whether non-believers separately or jointly process signals that arrive in separate clumps. If a non-believer naively anticipates separating new signals but after the fact infers by pooling all signals, he can be caught in a “learning trap”— persistently paying for information that he thinks will let him reach a decision, but which in retrospect he finds inconclusive. In observational-learning settings, non-believers can end up in an information cascade where they ignore their private information, but (unlike believers) may also end up in an “informational eddy,” forever following their own signals.

JEL Classification: B49, D03, D14, D83, G11

*We thank Herman Chernoff, Botond Koszegi, Don Moore, Ted O’Donoghue, Larry Phillips, and Daniel Read and seminar participants at UC Berkeley, Cornell University, Koc University, the LSE-WZB Conference on Behavioral IO, and the Behavioral Economics Annual Meeting for helpful comments. We are grateful to Ahmed Jaber, Greg Muenzen, Desmond Ong, Nathaniel Schorr, Dennis Shiraev, Josh Tasoff, Mike Urbancic, and Xiaoyu Xu for research assistance. Raymond thanks the University of Michigan School of Information’s Socio-Technical Infrastructure for Electronic Transactions Multidisciplinary Doctoral Fellowship funded by NSF IGERT grant #0654014 for financial support. E-mail: daniel.benjamin@gmail.com, rabin@econ.berkeley.edu, cbraymon@umich.edu.

1 Introduction

Most people understand that larger samples are more likely to reflect the proportions of a population than smaller ones. Following Tversky & Kahneman (1971), Rabin (2002) and Rabin & Vayanos (2010) model the notion that people believe in “the Law of Small Numbers,” and exaggerate how likely it is that small samples will reflect the underlying population. Yet evidence indicates that people *also* do not believe in the Law of Large Numbers: they believe that even in very large random samples, proportions might depart significantly from the overall population rate. This paper develops a formal model of such “non-belief in the Law of Large Numbers,” which we abbreviate by NBLLN.¹ We explore the model’s relationship to psychological evidence, its implications for inference, and some of its economic consequences. We identify a few alternative ways that NBLLN may manifest itself, and we confront and explore conceptual challenges with modeling NBLLN that are also likely to arise in modeling other biases in statistical reasoning.

Kahneman & Tversky (1972) find that subjects seem to think sample proportions reflect a “universal sampling distribution,” virtually neglecting sample size.² In doing so, subjects vastly exaggerate the probability of unbalanced ratios in large samples. For instance, independent of whether a fair coin is flipped 10, 100, or 1,000 times, the median subject thinks that the probability of getting between 45% and 55% heads is about .21, and of getting between 75% and 85% heads is about .05.³ These beliefs are close to right for the sample size of 10, where the correct answers are .25 and .04. But for the sample size of 1,000, there is over a .99 chance of between 45% and 55% heads and a negligible chance of between 75% and 85% heads.

In Section 2, we develop our model of non-belief in the Law of Large Numbers in a simple setting, where a person is trying to predict the distribution of—or make an inference from—a fixed sample size. Throughout, we refer to our modeled non-believer in the Law of Large Numbers as *Barney*, and compare his beliefs and behavior to a purely Bayesian information processor, *Tommy*.⁴ Tommy knows the likelihood of different sample distributions of an *i.i.d.* coin biased θ towards heads will be the “ θ -binomial distribution.” But Barney believes that large-sample proportions will be distributed according to a “ β -binomial distribution,” for some $\beta \in [0, 1]$ that itself is drawn from a distribution with mean θ . This model directly implies NBLLN: whereas Tommy knows that

¹NBLLN is pronounced letter by letter, said with the same emphasis and rhythm as “Ahmadinejad.”

²Kahneman & Tversky (1972) associated such sample-size neglect with other biases under the rubric of “representativeness,” which are a range of ways that people reason statistically using judgments of similarity between hypotheses and observed evidence. We relate our modeling approach to representativeness and sample-size neglect in Section 7.

³These numbers come from eyeballing histograms contained in the paper.

⁴“Tommy” is the conventional designation in the quasi-Bayesian literature to refer somebody who updates according to the dictums of the Reverend Thomas “Tommy” Bayes.

large samples will have proportions of heads very close to θ , Barney feels that the proportions in any given sample, no matter how large, might not be θ .⁵ Even short of the large-sample limit, the model generates some properties largely in line with both the existing empirical evidence and the basic psychology others have proposed for N BLLN. For instance, while the two have identical beliefs about sample sizes of 1, Barney’s beliefs about sample proportions will be a mean-preserving spread of Tommy’s for samples of two or more signals. Although it embeds *some* sensitivity to sample sizes, the model largely reflects the “universal sample distribution” intuition from Kahneman & Tversky (1972).

Section 2 then draws out the implications of our simple model for *inference*. We show that if Barney applies full Bayesian updating based on his wrong beliefs about the likelihood of different sample realizations, N BLLN implies under-inference: for any priors and any sample size, Barney’s expected posterior ratio on different hypotheses will be less extreme on average than Tommy’s. Most importantly, for any non-extreme proportions of signals—including the limit proportion generated by the true state—Barney always fails to become fully confident even after infinite data.

In Section 3, we illustrate one economic implication of N BLLN by demonstrating the important role it likely plays in monetary-risk attitudes. In particular, Barney believes that the risk associated with a large number of independent gambles is greater than it actually is. This magnifies aversion to repeated risks, whether that risk aversion is due to diminishing marginal utility of wealth or (more relevantly) reference-dependent risk attitudes. Because he does not realize that the chance of aggregate losses becomes negligible, Barney may refuse to accept even infinite repetitions of a small, better-than-fair gamble. This refusal contradicts any plausible preferences combined with belief in the Law of Large Numbers, and yet reflects both the observed choices, most notably in Benartzi & Thaler (1999), who demonstrate clearly the role of what we are calling N BLLN, and the basic psychology of attitudes toward independent risks.

Both Sections 2 and 3 analyze a model of N BLLN when there is a single, “given” sample that Barney will observe. Yet information does not always arrive in a single package of signals. A person may hear a series of individual reports from friends about their car experiences, while also reading large-sample statistics of car performance. If Barney pools each of his friends’ tales with the statistics from *Consumer Reports*, his inferences will be very different than if he separately

⁵Although this modeling gimmick is “as if” Barney is unsure that the rate is θ , true parameter uncertainty is not at all our interpretation. Instead, consistent with the underlying N BLLN psychology, we interpret it as Barney’s belief that even his certainty that the underlying rate is θ is not a guarantee that the proportion in very large samples will approximate θ . In the dynamic model we develop below, this interpretation will be much more than an aspiration to get the psychology right, but an integral part of the formal model. For instance, in situations where Barney must predict further signals after observing his first 100 signals, we assume his expected proportions are still θ , rather than being influenced by the first signals as the parameter-uncertainty interpretation would suggest.

updates his beliefs following each of his friend’s stories, and then treats the *Consumer Reports* data as one big sample. Such cases confront us with a conceptual challenge intrinsic to the very nature of NBLLN: because people under-infer worse for larger samples than smaller ones, they will infer differently if they lump observations together versus separately. A model of NBLLN must involve a theory of how Barney groups information as a function of how it is presented to him and other features of his decisionmaking environment. With little empirical research to guide us, in Section 4 we discuss and formalize various combinations of assumptions on how Barney “retrospectively groups” signals—how he interprets evidence once he sees it—and “prospectively groups” signals—how he predicts ahead of time he will interpret evidence he might observe in the future. Different combinations of assumptions may be warranted by different perceptual, framing, and decisionmaking environments. Of special interest is the possibility that Barney retrospectively groups signals differently than he prospectively anticipates he will. He may, for instance, plan to separately ask friends about their experiences, and prospectively focus on each conversation as if it is a separate signal; but then in retrospect, he may pool the conversations together as a large sample.

In Section 5, we explore Barney’s behavior in various economic environments involving learning and inference, all of which reflect the central fact that Barney fails to reach appropriately strong confidence based on extensive evidence. When Barney decides as he goes along whether to gather more information, like Tommy, he will plan to quit his costly search once he reaches some threshold of confidence. And, like Tommy, he may stop gathering signals very quickly if information is decisive. But because Barney tends to infer far less from signals than Tommy, when the initial signals are mixed, Barney may continue trying to learn even after many signals. Indeed, if Barney prospectively anticipates separately updating using each arriving signal but actually pools them retrospectively, Barney may become stuck in a “learning trap”: he persistently expects to soon be confident enough to stop experimenting, or buying information, but, because he never achieves the confidence he anticipates, continues his costly efforts forever. In such cases, variants of NBLLN predict not only that people will never figure out the truth when they rationally should, but that their efforts to learn may be enormously costly.

In Section 6 we explore possible implications of NBLLN in an environment where Barney learns from observing both his own information directly and also from the actions of others with private information. The most celebrated result in research on this topic is that a sequence of Tommys who each get independent information and observe the full sequence of actions by others who choose before them will surely end up in an “information cascade” in which each ignores his own signal and instead mimics the actions of those he observes. Without thoroughly addressing modeling issues

that arise in multi-person settings, we draw out some implications of NLLN for herding based on a plausible strategic extension of Barneyess. Barney’s conservatism in updating means herding is likely to be slower for a society of Barneys than for one of Tommys. But a qualitative difference also arises: although in all cases information cascades occur with positive probability, there are also parameter values where there is a positive probability that a sequence of Barneys will never enter a cascade. If early signals do not generate strong evidence in favor of one action, Barneys may end up in an “information eddy,” in which each agent always chooses the action corresponding to his own signal.

In Section 7, we discuss why our model might be more compelling than alternative possible explanations and modeling approaches—both fully rational and not fully rational—that might seem to accommodate the psychology evidence. Our model of course ignores other important departures from Bayesian inference—such as base-rate neglect and belief in the Law of Small Numbers—that seem separable from NLLN. But it also omits features—including the psychophysics of diminishing sensitivity, as well as unwillingness to hold or express extreme beliefs—that, as alternative sources of under-inference from large samples, are less separable. In Appendix A, in fact, we present a (complicated) formal model embedding some of these other errors along with NLLN. Guided by this formal model, in Appendix B we attempt to give a fairly exhaustive review of the empirical work on sampling predictions and inference. We believe this review makes clear that our model of NLLN is capturing a broad empirical reality. Although in many settings there are alternative reasons why a person may rationally disregard a large amount of evidence—e.g., because the evidence is less relevant given the person’s preferences than other, smaller-sample sources of information—our review documents NLLN only in settings where a Bayesian would fully attend to the evidence. But NLLN is not a phenomenon that has been widely embraced or emphasized by judgment researchers or behavioral economists. We surmise that this is largely because findings of under-inference have been associated with an interpretation called “conservatism” (e.g., Edwards, 1968)—namely, that people tend not to update their beliefs as strongly as Bayesian updating dictates—that does not mesh comfortably with base-rate neglect and other biases that often imply that people infer more strongly than Bayesian. In our view, summarizing people as overly conservative or not conservative enough is manifestly the wrong way to parse human judgment. By focusing on the concrete biases at play and highlighting the co-existence of NLLN with other biases, we hope to make clear that there is no contradiction. Indeed, as we discuss and illustrate below, NLLN is often a necessary “enabler” of other psychological biases, including biases where people over-infer from some information. For example, another bias is that people seem to draw excessively strong conclusions from “vivid” signals. But even if observing a traffic accident first-hand has the impact

of 100 data points on a person’s beliefs about the likelihood of an accident, such a vivid signal would affect beliefs little unless the person *also* drew too weak an inference from the newspaper’s report of the past year’s accident statistics due to NBLLN. As another example, Barney’s failure to learn in an experimentation setting may help explain why people fail to fully learn the extent of their own present bias even after observing their own behavior in many decisionmaking contexts over many years.

Because we think that NBLLN is likely to matter for economic theory, we strive in this paper to develop a model that is both specific in its implications and general in its applicability. We believe the theorems and other results we establish below—in general and in the specific economic applications we study—makes clear its potential to provide an improvement over full rationality models while replicating as much as possible the specificity of the full-rationality predictions. In the spirit of seeing our model as an iterative step towards improving the psychological realism of formal economic models, we highlight throughout the paper cases where we believe our assumptions are restrictive and fail to capture psychological intuitions or existing evidence, and we outline in Appendix A ways that the model can be combined with other psychological biases. Also in this spirit, we conclude Section 7 by briefly discussing an approach to generalizing our model to contexts where Barney observes non-binomial or non-*i.i.d.* data, which would be a crucial step to incorporating NBLLN more broadly into economic models.

2 The Single-Sample Model

Throughout the paper, we study a stylized setting where an agent observes a set of binary signals, each of which takes on a value of either a or b . Given a *rate* $\theta \in \Theta \equiv (0, 1)$, signals are generated by an *i.i.d.* process where the probability that any given signal is an a is equal to θ . Signals arrive in *clumps* of size N . We denote the set of possible ordered sets of signals of size $N \in \{1, 2, \dots\}$ by $S_N \equiv \{a, b\}^N$, and we denote an arbitrary clump (of size N) by $s \in S_N$.⁶ Let A_s denote the total number of a ’s that occur in the clump $s \in S_N$, so that $\frac{A_s}{N}$ is the proportion of a ’s that occur in a clump of N signals. For a real number x , we will use the standard notations “ $\lceil x \rceil$ ” to signify the smallest integer that is weakly greater than or equal to x and “ $\lfloor x \rfloor$ ” to signify the largest integer that is weakly less than or equal to than x .⁷ For any random variable y that takes as possible values the elements of set Y , let beliefs by Tommy (the perfect Bayesian information processor)

⁶Note that we forego the conventional strategy of providing notation for a generic signal, indexed by its number. It is less useful here because (within a clump) what matters to Barney is just the number of a signals, not their order. Conserving notation here, in Section 4 we use t to index the clumps of signals.

⁷Although our formal results and proofs fully attend to “integer issues,” we ignore them when providing intuition.

be denoted by probability density function $f_Y(\cdot)$, implying cumulative distribution function $F_Y(\cdot)$, expectation $E_Y(\cdot)$, and variance $Var_Y(\cdot)$. Let corresponding beliefs by Barney (the non-believer in the LLN) be denoted by $f_Y^\psi(\cdot)$, $F_Y^\psi(\cdot)$, $E_Y^\psi(\cdot)$, and $Var_Y^\psi(\cdot)$, where ψ is a parameter for the degree of NBLLN defined below.

In this section, we develop our model of Barney for the case where he is considering a single clump of N signals. This case corresponds to most of the experimental evidence about NBLLN, which has been collected in settings where subjects were presented with a single, fixed sample of signals or outcomes, where subjects presumably process all the information together. This special case also allows us to lay bare the essential features of how our model captures NBLLN. When generalizing the model, complicating conceptual challenges arise about which there is little existing evidence. Some of our analysis in fact concerns precisely these complications, but we defer discussion of these issues and ways to handle them until Section 4.

According to the Law of Large Numbers, the mean of a random sample equals the rate with probability 1 in the limit as the sample size gets large: For any interval $[\alpha_1, \alpha_2] \subseteq [0, 1]$,

$$\lim_{N \rightarrow \infty} \sum_{x=\lceil \alpha_1 N \rceil}^{\lceil \alpha_2 N \rceil} f_{S_N | \Theta}(A_s = x | \theta) = \begin{cases} 1 & \text{if } \theta \in [\alpha_1, \alpha_2] \\ 0 & \text{otherwise} \end{cases}.$$

How might we capture the possibility that Barney believes (say) that it is reasonably likely that at least 600 of 1000 births at a hospital in a given year are boys, even though he knows that boys are born at a rate of 50%? Our key modeling gambit is to assume that Barney believes samples are generated as if a rate of θ , here 50%, means that the rate is θ on average, but might be higher or lower for any given sample. For a given true rate θ , we model Barney as believing that for the sample he is considering: first, a “subjective rate” $\beta \in [0, 1]$ is drawn from a distribution centered at θ . Then the *i.i.d.* sample of 1000 babies is generated using rate β . The key implication is that if a given value of β were the actual rate, it would (by the Law of Large Numbers!) exactly determine the proportion of signals in the limit of a very large sample. Therefore, the probability density that Barney assigns to any proportion β of signals (say, 60% of babies are boys) in a large sample is equal to the probability density that Barney assigns to the possibility that β equals that value.

Formally, we assume that when Barney knows the rate is θ , he believes that signals are generated by an *i.i.d.* process with *subjective rate* $\beta \in [0, 1]$, where β is drawn from a density $f_{\beta | \Theta}^\psi(\beta | \theta)$ —which we refer to as Barney’s *subjective rate distribution*—that has the following properties.

A1. For all β and θ , $f_{\beta | \Theta}^\psi(\beta | \theta)$ has full support on $(0, 1)$, is absolutely continuous in β , and is point-wise continuous in θ .

A2. For all β and θ , $F_{\beta|\Theta}^\psi(\beta|1-\theta) = 1 - F_{\beta|\Theta}^\psi(1-\beta|\theta)$.

A3. For all θ and $\theta' > \theta$, $\frac{f_{\beta|\Theta}^\psi(\beta|\theta')}{f_{\beta|\Theta}^\psi(\beta|\theta)}$ is increasing in β .

A4. For all β and θ , $E_{\beta|\Theta}^\psi(\beta|\theta) = \theta$.

Assumptions A1 and A2 are mild and consistent with the empirically-observed densities of people’s beliefs about the distribution of signals in large samples. Assumptions A3 and A4, however, are substantive assumptions that do not follow easily from the psychology. Assumption A3 is a monotone likelihood ratio property: fixing any two rates, Barney believes that the likelihood of drawing any particular subjective rate given the high rate relative to the low rate is greater the higher is the high rate. This is a substantive assumption,⁸ and it is easy to imagine specifications of $f_{\beta|\Theta}^\psi(\beta|\theta)$ —especially in the spirit of the type of diminishing-sensitivity evidence discussed in Appendix A—that would violate it. But it is in accord with the most directly relevant evidence, namely Griffin & Tversky’s (1992) Study 3, which examines a range of parameters of the sort that seems most likely for violating A3.⁹ It holds for our main example of the beta distribution (discussed below) and more generally is useful for establishing some of our results. Especially because the range of samples for which it is potentially false are inherently very unlikely, we think it is probably not an important caveat to our results. Assumption A4 says that the mean of Barney’s subjective rate distribution is the known objective rate. Although we rely on it extensively in the analysis, it is in fact violated in existing data; we discuss below both the violations and the role it is playing in our analysis.

When Barney knows the rate is θ , he believes the likelihood of observing a particular clump of N signals, $s \in S_N$, is

$$f_{S_N|\Theta}^\psi(s|\theta) = \int_{\beta \in [0,1]} f_{S_N|\beta}(s|\beta) f_{\beta|\Theta}^\psi(\beta|\theta) d\beta, \quad (1)$$

where $f_{S_N|\beta}(s|\beta)$ is the (correct) probability of observing s if the rate were β , and this is averaged over the density of subjective rates, $f_{\beta|\Theta}^\psi(\beta|\theta)$. Consequently, Barney’s belief that a large sample

⁸For example, Whitt (1979) shows that this monotone likelihood ratio property implies: for all β , θ , and $\theta' > \theta$, $F_{\beta|\Theta}^\psi(\beta|\theta')$ first-order stochastically dominates (FOSD) $F_{\beta|\Theta}^\psi(\beta|\theta)$. This FOSD relationship is itself useful in proving some of our results.

⁹In particular, Griffin & Tversky asked subjects to infer the likelihood that a coin is biased $\theta_A = .6$ in favor of heads rather than $\theta_B = .25$ in favor of heads, depending on different possible outcomes from flipping the coin 12 times. There is a psychological (but statistically erroneous) intuition that extreme samples, like 10 heads out of 12, seem so unexpected in the case of either rate that they do not provide strong evidence about which rate is generating the flips. This diminishing-sensitivity intuition might lead people to think that the evidence in favor of the .6-biased coin is stronger when the observed sample is 7 heads out of 12 than when it is 10 heads out of 12—in violation of the implications of Assumption A3 for inference. However, consistent with A3, Griffin & Tversky find that subjects’ posteriors beliefs in favor of the .6-biased coin are monotonically increasing in the number of heads.

will have a proportion of a signals in some range $[\alpha_1, \alpha_2]$ is exactly equal to Barney’s belief that the subjective rate β is in that range.¹⁰

Lemma 1. *Barney does not believe in LLN: for any interval $[\alpha_1, \alpha_2] \subseteq [0, 1]$,*

$$\lim_{N \rightarrow \infty} \sum_{x=[\alpha_1 N]}^{[\alpha_2 N]} f_{S_N|\Theta}^\psi(A_s = x|\theta) = F_{\beta|\Theta}^\psi(\beta = \alpha_2|\theta) - F_{\beta|\Theta}^\psi(\beta = \alpha_1|\theta) > 0.$$

Because we assume that Barney’s beliefs about the distribution of β puts positive probability density on the entire interval $(0, 1)$, the subjective-rate modeling gimmick captures the essence of our interpretation of NLLN: Barney believes that the proportion of heads from flipping a coin *known to be fair* may not be 50% in any given sample, no matter how large. In keeping with this interpretation, Barney does not believe the realized β is a real feature of the coin, and is certainly not an object he makes inferences about. It is a representation of Barney’s subjective uncertainty that the θ will manifest itself in a given sample. In the analysis of Barney’s predictions about a single clump of signals, our model is mathematically equivalent to a fully rational model where the agent is uncertain about the underlying rate. But both the inference results later in this section and Section 4’s dynamic setting demonstrates how different these two interpretations are: if an agent were rationally uncertain about the underlying rate, then in large samples, he would (quickly) become certain about its value. In contrast, NLLN means that the agent acts as if he is uncertain about the underlying rate even when he knows what the rate is or has seen enough data that he should be sure about it.

Since (by the Law of Large Numbers) Barney’s belief about the distribution of signals in large samples coincides with his subjective rate distribution, the most appropriate density of $\beta|\theta$ would correspond to the empirical beliefs in studies such as those illustrated in Figure 1, drawn from Kahneman & Tversky (1972). The black, gray, and white bars—which correspond to people’s reported beliefs regarding samples of size 10, 100, and 1000, respectively—lie on top of each other. This distribution that is common across sample sizes, which presumably corresponds to people’s large-sample beliefs about sample proportions, could be directly assumed to be the density of $\beta|\theta$.

Although Assumptions A1 and A2 are consistent with Figure 1, A4 is not. Beliefs for $\theta = .5$, depicted in the left panel, naturally have mean approximately equal to .5. However, beliefs for $\theta = .8$, depicted in the right panel, have mean approximately equal to .6. The mean of the distribution of signals is displaced toward .5 apparently because the long tail of the distribution is

¹⁰All proofs are elevated to Appendix C.

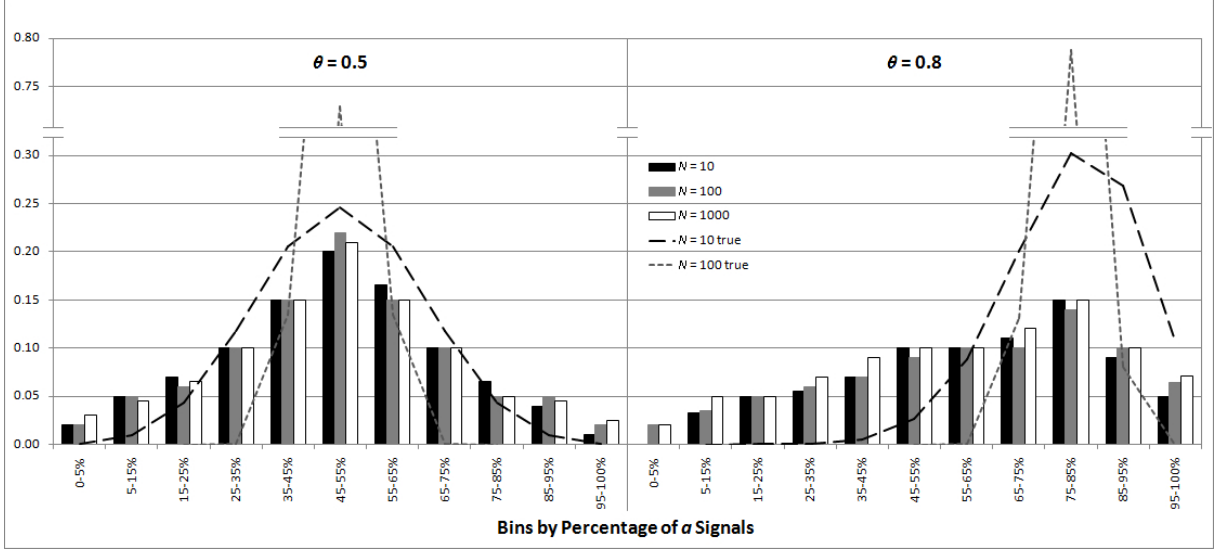


Figure 1: Evidence from Kahneman & Tversky (1972)

fat. As we discuss in Appendices A and B, we believe that the fatness of the tail is in turn due to *flatness* of the tail. We attribute the flatness to a psychological bias—“sampling-distribution-tails diminishing sensitivity”—in which people perceive very unlikely outcomes as similar to each other and hence similar in probability. We justify A4 on two grounds: analytical convenience, and our contention that the violation of the assumption is best understood as a distinct psychological bias modeled in Appendix A. Nonetheless, because we omit flat tails from the model, the model will not match some features of the empirical evidence, especially when the agent observes an extreme sample.¹¹

A *subjective sampling distribution* specifies an agent’s belief about the likelihood of each possible combination of signals when the rate θ is known. Whereas Lemma 1 shows that Barney’s subjective sampling distribution (for the number of a -signals) in the large-sample limit equals his “subjective rate distribution,” Proposition 1 shows that NBLLN also has implications for Barney’s finite-sample subjective sampling distributions.

Proposition 1. *For any $\theta \in (0, 1)$ and $N \in \{1, 2, \dots\}$:*

1. $E_{S_N|\Theta}^\psi \left(\frac{A_s}{N} | \theta \right) = E_{S_N|\Theta} \left(\frac{A_s}{N} | \theta \right) = \theta$.
2. $F_{S_N|\Theta}^\psi (A_s | \theta)$ *second-order stochastically dominates (SOSD)* $F_{S_N|\Theta} (A_s | \theta)$, and $\text{Var}_{S_N|\Theta}^\psi \left(\frac{A_s}{N} | \theta \right) \geq$

¹¹Analogously, we justify omitting the Law of Small Numbers (LSN) from the model for analytical convenience, and our contention that it is a distinct bias. However, our grounds for not incorporating LSN into the model are stronger because LSN will matter most when NBLLN will matter least—when the observed sample is small—whereas sampling-distribution-tails diminishing sensitivity is, like NBLLN, an especially extreme bias when the observed sample is large.

$Var_{S_N|\Theta} \left(\frac{A_s}{N} | \theta \right)$ with strict inequality for $N > 1$.

3. $Var_{S_N|\Theta} \left(\frac{A_s}{N} | \theta \right)$ is strictly decreasing in N , and $\lim_{N \rightarrow \infty} Var_{S_N|\Theta} \left(\frac{A_s}{N} | \theta \right) > 0$.
4. $F_{S_N|\Theta}^\psi (A_s | \theta')$ first-order stochastically dominates (FOSD) $F_{S_N|\Theta}^\psi (A_s | \theta)$ whenever $\theta' > \theta$.

Part 1 states that Barney, like Tommy, expects the average proportion of a 's in the sample to be θ . An immediate and important corollary of Part 1 is that Barney's beliefs coincide with Tommy's when $N = 1$. Part 2 states that Barney has a more-dispersed subjective sampling distribution than Tommy in the sense of second-order stochastic dominance. Combined with the fact that the mean of Barney's subjective sampling distribution is the same as Tommy's, this implies that Barney's subjective sampling distribution is a mean-preserving spread of Tommy's. This naturally implies that the variance of Barney's subjective sampling distribution is larger than Tommy's. Part 3 states that although the variance of Barney's subjective sampling distribution (for the sample proportion) is strictly decreasing in N , it is bounded away from 0. Part 4 states that a higher true rate, besides generating a higher mean proportion of a signals, generates a rightward shift in Barney's entire subjective sampling distribution in the sense of first-order stochastic dominance.

We now turn to *inference* problems, where an agent with prior beliefs must infer from observed signals what the underlying rate is. An example of an inference problem is determining the likelihood that a coin is head-biased rather than tail-biased, after observing a sample of coin flips. Let $\Theta \subseteq (0, 1)$ denote the set of rates that have positive prior probability. For simplicity, we assume Θ is a finite set. Without loss of generality, we consider the agent's beliefs about the relative likelihood of two of the rates $\theta_A > \theta_B$, given priors $0 < f_\Theta(\theta_A), f_\Theta(\theta_B) < 1$ with $f_\Theta(\theta_A) + f_\Theta(\theta_B) \leq 1$.

We maintain the standard assumption that an agent draws inferences by applying Bayes' Rule to his subjective sampling distributions.¹² Consequently, Barney's beliefs after observing a particular clump $s \in S_N$ are $f_{\Theta|S_N}^\psi(\theta_A | s) = \frac{f_{S_N|\Theta}^\psi(s|\theta_A)f_\Theta(\theta_A)}{\sum_{\theta \in \Theta} f_{S_N|\Theta}^\psi(s|\theta)f_\Theta(\theta)}$ and $f_{\Theta|S_N}^\psi(\theta_B | s) = \frac{f_{S_N|\Theta}^\psi(s|\theta_B)f_\Theta(\theta_B)}{\sum_{\theta \in \Theta} f_{S_N|\Theta}^\psi(s|\theta)f_\Theta(\theta)}$. Because Barney updates Bayesianly given his model of the data-generating process, his subjective beliefs satisfy the "Law of Iterated Expectations": Barney expects that for any sample size, the mean of his posterior beliefs will equal the mean of his prior beliefs: for any $N \geq 1$, $E_{\Theta|S_N}^\psi(\theta | s) = E_\Theta^\psi(\theta)$.

Due to the LLN, after observing a sufficiently large number of signals, Tommy will be arbitrarily close to knowing the true rate with certainty. In contrast, the central implication for inference of

¹²Appendix A discusses and formalizes how to combine NLLN with an important way people do not draw inferences in accordance with Bayes' Rule: they underweight priors according to the well-established notion of base-rate neglect. In the main model, however, we maintain the standard assumption of Bayesian inference both to highlight the role played *per se* by NLLN, and because (as we discuss in Appendix B) our reading of the experimental evidence is that *except* for base-rate neglect, people's inferences are actually well-approximated by Bayes' Rule applied to their subjective sampling distributions. Hence in applications where priors are equal, base-rate neglect is neutralized as a factor. In applications where we do not assume equal priors, however, base-rate neglect could modify some of our results.

Barney’s NBLLN—which plays a large role in many of the applications later in this paper—is that Barney remains uncertain even after observing an *infinite* number of signals. To boot:

Proposition 2. *Let $\theta \in (0, 1)$ be the true rate. Then for any $\theta_A, \theta_B \in \Theta$, Barney draws limited inference even from an infinite sample: as $N \rightarrow \infty$,*

$$\frac{f_{\Theta|S_N}^{\psi}(\theta_A|s)}{f_{\Theta|S_N}^{\psi}(\theta_B|s)} \xrightarrow{a.s.} \frac{f_{B|\Theta}^{\psi}(\beta = \theta|\theta_A) f(\theta_A)}{f_{B|\Theta}^{\psi}(\beta = \theta|\theta_B) f(\theta_B)}, \quad (2)$$

which is a positive, finite number.

Because Barney’s asymptotic sampling distribution coincides with the subjective-rate distribution, his limit inference depends on the relative heights that the pdfs of the subjective-rate distributions for θ_A and θ_B assign to the proportion θ of a ’s. Since the subjective-rate distributions put positive density on every proportion in $(0, 1)$, Barney’s likelihood ratio will be finite. An immediate but important corollary of Proposition 2 is that Barney’s priors affect his beliefs even in the limit of an infinite sample.

Proposition 2 shows that, like Tommy’s, Barney’s limit posterior is a deterministic function of the true rate, but unlike Tommy’s, Barney’s limit posterior ratio is bounded away from 0 and ∞ . Not only will Tommy learn the true rate for sure after observing a sufficiently large number of signals, Tommy also correctly *anticipates* that a sufficiently large number of signals will make him certain of the true rate. In contrast, Barney does not even realize that his limit posterior is a deterministic function of the true rate. Barney mistakenly thinks that his posterior probability of rate θ_A after observing an infinite number of signals is a random function of the true rate. The reason is that, even though Barney knows that his inferences in a large sample will be determined by the proportion of a ’s, he incorrectly thinks the proportion of a ’s is determined by the subjective rate, which is a random function of the true rate.

Proposition 3. *Fix any $\theta_A > \theta_B \in \Theta$. Barney believes that, regardless of what the true rate $\theta \in (0, 1)$ is, as $N \rightarrow \infty$, his posterior ratio $\frac{f_{\Theta|S_N}^{\psi}(\theta_A|s)}{f_{\Theta|S_N}^{\psi}(\theta_B|s)}$ will converge in distribution to a random variable whose support is a simply connected subset of $[0, \infty]$. Moreover, he believes that this subset of $[0, \infty]$ has strictly positive measure.*

The upper bound on Barney’s anticipated limit posterior corresponds to what Barney’s limit posterior would be if he observed all a -signals; the lower bound corresponds to his limit posterior if he observed all b -signals.

Because Barney’s subjective sampling distribution is correct when the sample size is 1, he will draw correct inferences in that case. Barney will also draw correct inferences when exactly half the signals are a -signals. In that case, the sample is uninformative for both Barney and Tommy, and neither updates his beliefs about the rate.

Proposition 4. *Fix rates $\theta_A > \theta_B$. For $N = 1$, Barney and Tommy infer the same. If $\theta_A = 1 - \theta_B$, then for any set of $N \in \{1, 2, \dots\}$ signal realizations $s \in S_N$, neither Tommy’s beliefs nor Barney’s beliefs change from the priors when $\frac{A_s}{N} = \frac{1}{2}$.*

For further analysis of the finite-sample case—as well as for theoretical applications and empirical analysis—it is useful to have a parametric model of Barney’s subjective rate distribution. For some of our results, we impose a particular functional form for the subjective rate distribution, a beta distribution:

$$f_{\beta|\Theta}^{\psi}(\beta|\theta) = \beta^{\theta\psi-1} (1 - \beta)^{(1-\theta)\psi-1} \frac{\Gamma(\psi)}{\Gamma(\theta\psi)\Gamma((1-\theta)\psi)}, \quad (3)$$

where $0 < \psi < \infty$ is the exogenous parameter of the model, and $\Gamma(x) \equiv \int_{[0,\infty)} y^{x-1} e^{-y} dy$, defined on $x > 0$.¹³ The properties A1-A4 are satisfied, and this family of beta densities shares many qualitative features of people’s empirically observed large-sample beliefs about the distribution of signals.¹⁴ A major advantage of this formulation is tractability: since the beta distribution is the conjugate prior for the binomial distribution, standard results from probability theory can be used to characterize Barney’s beliefs.¹⁵ “Parameterized-Barney” is more biased for smaller ψ —with

¹³The more common way of writing this beta density is $f_{\beta|\Theta}^{\psi}(\beta|\theta) = \beta^{\theta\psi-1} (1 - \beta)^{(1-\theta)\psi-1} \frac{(\psi-1)!}{(\theta\psi-1)!((1-\theta)\psi-1)!}$. Our formulation is equivalent, except it allows for non-integer values of ψ . Recall that the Gamma Function, $\Gamma(x)$, is the standard generalization of the factorial function: it has the properties that $\Gamma(x+1) = x\Gamma(x)$ and $\Gamma(1) = 1$, so that for any positive integer x , $\Gamma(x) = (x-1)!$.

¹⁴We prove that properties A1-A4 are satisfied in Lemma $\beta 5$ in Appendix C. Although we do not assume it in A1-A4, the available empirical evidence is also consistent with the subjective rate distribution being single peaked. Single peakedness, however, is not satisfied for all parameter values of the beta distribution. If $\theta\psi$ and $(1-\theta)\psi$ are both less than 1, the beta distribution has a first derivative with respect to β that is increasing everywhere—the pdf has a minimum in the interior of $(0, 1)$ and rises as it approaches both 0 and 1.

¹⁵Unfortunately, the functional form (3) has a few implications about asymmetric inference (i.e., inference problems where $\theta_A \neq 1 - \theta_B$) that can matter in applications but that do not have general intuitions related to NBLLN: (1) In asymmetric-inference problems, parameterized Barney draws the same inferences as Tommy when the observed sequence is ab or ba ; (2) In asymmetric-inference problems, when the sample is all a ’s or all b ’s, parameterized Barney always under-infers regardless of the specific values of θ_A and θ_B ; and (3) For any true rate $\theta \in (0, 1)$ and for any $\theta_A, \theta_B \in \Theta$, Barney believes that his limit posterior ratio, $\lim_{N \rightarrow \infty} \frac{f_{\Theta|S_N}^{\psi}(\theta_A|s)}{f_{\Theta|S_N}^{\psi}(\theta_B|s)}$, will be a random variable that has support on all of $(0, \infty)$. These claims are proved in

Lemma $\beta 4$ in Appendix C. Since we believe these properties would not generalize to other models that are equally consistent with existing evidence, we will avoid stating implications that rely on these properties of the functional form.

more dispersed subjective sampling distributions in the sense of SOSD—and Barney coincides with Tommy in the parameter limit $\psi \rightarrow \infty$. To interpret the “Barneyness parameter,” ψ , consider the problem of inferring the rate from a sample of signals. The beta density (3) corresponds to what a Bayesian’s posterior about the rate would be if the Bayesian had begun with an improper beta prior on $[0, 1]$ corresponding to the $\psi \rightarrow 0$ limit of (3), and had observed a total of ψ signals with $\theta\psi$ a ’s and $(1 - \theta)\psi$ b ’s. For $\psi = 0$, the posterior would remain the improper prior on $[0, 1]$, while for $\psi \rightarrow \infty$, the posterior would converge to a point mass at the true rate θ . Although we do not conduct a full and careful structural estimation, using evidence from studies that elicit subjects’ subjective sampling distribution as well as those that measure subjects’ inference about rates after observing signal realizations, we estimate that ψ falls within a range of 7-15. To illustrate how powerful a bias NBLLN is when $\psi = 10$, it implies that Barney’s beliefs about whether a coin is a 40% heads coin or a 60% heads coin are the same after an infinite number of signals of 60% heads as Tommy’s are after only 6 heads and 4 tails.

This parameterized model of Barney gives a sense of magnitudes for how Barney’s under-inference depends on the rates θ_A and θ_B . Suppose $\psi = 10$, Barney begins with equal priors on the two states, and the true rate is θ_A . If the difference between the rates is relatively large—with $\theta_A = 1 - \theta_B = .8$ —then the role of NBLLN is relatively small. In an infinite sample, Barney’s subjective posterior probability of rate θ_A will converge to .9998. However, if the two rates are closer together—with $\theta_A = 1 - \theta_B = .6$ —then in an infinite sample, Barney’s subjective posterior probability of rate θ_A will converge to only .69. As a reminder about the role of priors, this in fact means if Barney initially had beliefs more extreme than 2.25:1 in favor of rate θ_B , he will, in an infinite sample, surely end up believing rate θ_B is more likely, even when θ_A is the true rate.

While most dramatic in large samples, NBLLN also has implications for inference in finite samples. Since Barney’s subjective sampling distribution is too dispersed when $N > 1$, Barney will “generally” under-infer when the sample size is larger than 1. In order to make that claim precise, we will measure Barney’s (and Tommy’s) “change in beliefs” by the absolute difference between his posterior probability that θ_A is the true rate and his prior probability: $\left| f_{\Theta|S_N}^{\psi}(\theta_A|s) - f_{\Theta}(\theta_A) \right|$. We say that Barney “under-infers” relative to Tommy if Barney’s change in beliefs is smaller than Tommy’s.

Unlike in large samples, in small samples it is no longer universally true that Barney under-infers relative to Tommy. For particular realizations Barney can over-infer or under-infer relative to Tommy—or even infer in the opposite direction, so that a sample that causes Tommy to think rate θ_A is more likely, causes Barney to think rate θ_B is more likely!¹⁶ Nonetheless, we believe

¹⁶For example, using the parameterized model discussed below, set $\psi = 10$, $\theta_A = .7$, and $\theta_B = .6$, and

that Barney under-infers *in expectation*, taken with respect to the true sampling distribution. Proposition 5 proves this statement for the case of ψ sufficiently small, but we conjecture that it holds for any $0 < \psi < \infty$.¹⁷

Proposition 5. *Fix rates $\theta_A > \theta_B$ and a set of $N \in \{1, 2, \dots\}$ signal realizations $s \in S_N$. If $f_{B|\Theta}^\psi(\beta|\theta)$ has the functional form (3), then regardless of whether the true rate is θ_A or θ_B , for ψ sufficiently small, the expected change in Barney’s beliefs is smaller than the expected change in Tommy’s beliefs. Furthermore, suppose $\theta_A = 1 - \theta_B$. Then for any sample of $N > 1$ signals such that $\frac{A_s}{N} \neq \frac{1}{2}$ and any ψ , Barney under-infers relative to Tommy. In addition, while Tommy’s inference depends solely on the difference in the number of a and b signals, Barney’s change in beliefs is smaller from larger samples with the same difference.*

Intuitively, on average Barney under-infers because he partially attributes the information in the realized sample to the subjective rate, rather than extracting all of the information about the true rate.

In many of the inference experiments reviewed in Appendix B and in many of our applications involving inference, the two rates are “symmetric” in the sense that $\theta_A = 1 - \theta_B$, e.g., an urn might have either 60% red balls or 40% red balls. In this case, Proposition 5 shows that stronger comparisons can be made between Barney and Tommy: as long as the realized sample is informative, parameterized-Barney will under-infer, not just in expectation. Proposition 5 also notes a key feature of Barney’s updating that shows how it leads to a bias towards “proportional thinking” in inference along the lines suggested by researchers such as Griffin & Tversky (1992). Consider samples where the difference between the number of a -signals and the number of b -signals is the same, e.g., aa and $abababaa$. Tommy will draw the same inference from the two samples. But because his asymptotic sampling distributions depend on the proportion of a and b signals rather than their number, Barney infers less from the larger sample.

3 Application: Perceived Aggregate Risk

An economics professor at MIT once told his colleague Paul Samuelson that, whereas he would reject a bet for even odds to gain \$200 or lose \$100, he would accept 100 repetitions of that

assume equal priors on the two states. Then if the realizations of 80 signals are 53 a -signals and 27 b -signals, then Tommy believes that state A is more likely, while Barney believes state B is more likely.

¹⁷We have simulated Barney’s and Tommy’s expected change in beliefs for a range of parameter values: for each of $\psi \in \{1, 2, \dots, 30\}$ and $N \in \{5, 10, 15, 20\}$, we examined each of $\theta_A, \theta_B \in \{.5, .6, .7, .8, .9\}$. We also ran a number of simulations for $\theta_A = .99$ and $.999$ and for $\psi = 100$. In every case we examined, Barney’s expected change in beliefs was smaller than Tommy’s.

bet. Even though such behavior sounds reasonable to most of us, Samuelson (1963) reports the conversation so as to prove that it violates classical expected-utility theory. That is, a Tommy with classical expected-utility preferences (defined over final wealth) who does not exhibit unrealistically large wealth effects should be willing to take a single bet if and only if he is willing to take $N \geq 1$ independent plays of that bet. Intuitively, in the absence of wealth effects, preferring $K + 1$ bets to K bets is the same thing as preferring 1 bet on top of any realization of the K bets. By induction, preferring to take any positive number of the bets is the same as preferring to take one bet.

Yet, it is not just the “switching” that violates classical expected utility preferences, but the aversion to the single bet to begin with. Rabin (2000a, 2000b) and Rabin & Thaler (2001) have followed others in noting that the degree of concavity required for classical expected-utility preferences, defined over wealth, to explain risk-averse behavior over the small stakes involved in a single play of the gamble is calibrationally implausible. Loss aversion—the tendency to feel a loss more intensely than an equal-sized gain—explains why the majority of people who turn down the one-shot gamble do so.¹⁸ A Tommy with a simple, piecewise-linear loss-averse utility function

$$u(w_0, z) = \begin{cases} w_0 + z & \text{if } z \geq 0 \\ w_0 + \lambda z & \text{if } z < 0 \end{cases}, \quad (4)$$

where w_0 is initial wealth and z is a monetary gain or loss, will refuse the one-shot bet as long as the coefficient of loss aversion, λ —typically estimated to be around 2.25 (e.g., Tversky and Kahneman (1991))—is greater than 2.¹⁹ However, a Tommy with typical loss-averse preferences would be extremely happy to accept 100 repetitions of the same gamble: while the expected gain is \$5,000, the chance of a net loss is only 1/700, and the chance of losing more than \$1,000 is only 1/26,000.

Despite being loss averse enough to turn down the one bet, nobody with fully rational beliefs would turn down 100 repetitions of this bet. Yet, unlike Samuelson’s colleague, many people would do so! In hypothetical questions from one study in Benartzi & Thaler (1999), for instance, 36% of participants said they would turn down a single scaled-down Samuelson type bet (win \$100 or lose \$50); but fully 25% also reject the 100-times repeated gamble.²⁰ NBLLN helps explain why

¹⁸Samuelson himself had speculated that it was the willingness to accept repeated plays of the bet that was the mistake, rather than the refusal to accept a single gamble. Samuelson’s conjecture that his colleague’s willingness to accept the repeated gamble was the result of a “fallacy of large numbers”—a mistaken belief that the riskiness of the gamble evaporates with a sufficiently large number of repetitions—is the *opposite* of NBLLN, and is contradicted by Benartzi and Thaler’s (1999) evidence, reported below, that people exaggerate the probability of a loss in the repeated bet.

¹⁹Although essentially correct for small gambles, assuming linear consumption utility can become problematic if bets are repeated so many times as to involve large amounts of wealth. However, if in our limit results below, we halve the stakes every time we double the number of repetitions, the linearity assumption is unobjectionable.

²⁰In two other subject pools, they find 34% and 23% turn down a simple \$20/\$10 gamble, and *more*

many people turn down these gambles.²¹ Barney exaggerates the probability that the repeated bet will turn out badly. Indeed, Benartzi & Thaler report evidence that NBLLN is implicated in these choices: when asked the probability of losing money after 150 repetitions of a 90%/10% bet to gain \$0.10/lose \$0.50, 81% percent of subjects overestimated the probability—and by an enormous margin. While the correct answer is .003, the average estimate was .24.²² To show that subjects’ mistaken beliefs were driving their choices, Benartzi & Thaler compared subjects’ willingness to accept the repeated bet with their willingness to accept a single-play bet that had the histogram of money outcomes implied by the repeated bet. While only 49% of the college-student subjects accepted 150 repetitions of the bet, 90% accepted the equivalent single-play bet, suggesting that the repeated bet would have been very attractive if subjects had correctly understood the distribution of outcomes.^{23,24,25}

Formally, while a loss-averse Tommy will always accept a better-than-fair bet if it is repeated enough times, a loss-averse Barney may—depending on how favorable the bet is and how loss-averse he is—turn down an infinitely-repeated bet.

Proposition 6. *Suppose Barney and Tommy have simple, piecewise-linear loss-averse preferences*

people—57% and 50%—turn down the repeated gamble. Keren (1991) finds similar results in incentivized single bets vs five-times-repeated bets; for related hypothetical evidence, see Keren & Wagenaar (1987) and Redelmeier & Tversky (1992).

²¹Our emphasis on how NBLLN helps explain why loss-averse people turn down the repeated bet is because of its calibrational relevance, but it may be worth noting that NBLLN also has implications for how expected-utility-over-wealth agents respond to repetitions of bets. We can extend the “if” part of Samuelson’s theorem: if Barney rejects a bet at all initial wealth levels w_0 , then he would also reject any $N \geq 1$ independent plays of that bet. The “only if” direction does not extend, and a Barney who is just indifferent between accepting and rejecting a simple bet would, because he exaggerates the risk, strictly prefer to reject repeated versions of the gamble.

²²Results were essentially identical when subjects were asked the mirror-image problem, the probability gaining money after 150 repetitions of a 10%/90% bet to gain \$0.50/lose \$0.10.

²³Note also that something more than the type of “narrow bracketing” stressed by authors such as Tversky & Kahneman (1986), Kahneman & Lovallo (1993), Benartzi & Thaler (1995), Read, Loewenstein & Rabin (1999), Barberis, Huang & Thaler (2006), and Rabin & Weizsäcker (2009) seems to be playing a role. Those papers emphasize that people often react to a combination of risky bets as if they were deciding about each risky bet in isolation from all the others. While such neglect of the effects of aggregating risks may help explain why people reject the repeated gamble, it seems clear that even people who attend to the aggregate effects *misunderstand* these aggregate effects. Benartzi & Thaler (1999) make this especially clear by demonstrating directly that people asked the probability of aggregate loss of independent bets exaggerate along the lines predicted by NBLLN.

²⁴Benartzi & Thaler also elicited the effects of showing the histogram in the above hypothetical examples and showed that it reduces rejections from 25%, 57%, and 50% to, respectively, 14%, 10%, and 17%.

²⁵Klos, Weber & Weber (2005) replicate and extend Benartzi & Thaler’s findings. Klos, Weber & Weber present subjects with four lotteries, each of which may be played singly, repeated 5 times, or repeated 50 times. Subjects generally prefer the repeated gambles and correctly understand that the variance of monetary outcomes increases with number of repetitions, but in the repeated gambles, subjects vastly overestimate the probability of loss as well as the expected loss conditional on losing money. Klos, Weber & Weber also find that, even though the probability of the monetary outcome falling within a given interval around the expected value falls with the number of repetitions, subjects incorrectly believe it increases. This last finding is inconsistent with our model of NBLLN and may reflect a bias from focusing subjects’ attention on the expected value, or it may be consistent with “exact representativeness,” a bias we discuss in Appendix B.

as specified in (4). Fix any gamble (θ, h, t) , paying off $h > 0$ with probability θ and $-t$ with probability $1 - \theta$, that is better than fair: $\theta h > (1 - \theta)t$. For any $\lambda \geq 1$, there is some $N' \geq 1$ such that if $N > N'$, then Tommy will accept N repetitions of the gamble. In contrast, for Barney there is some threshold level of loss aversion $\hat{\lambda} > 1$ such that: if $\lambda < \hat{\lambda}$, then there is some N' sufficiently large such that Barney will accept N repetitions of the gamble for all $N > N'$; and if $\lambda \geq \hat{\lambda}$, then there is some N'' sufficiently large such that Barney will reject N repetitions of the gamble for all $N > N''$.

Moreover, it is possible for Barney to exhibit the *opposite* pattern from Samuelson's colleague, accepting the single bet but rejecting the 100-times repeated bet! Consistent with this possibility, the evidence cited above from Benartzi and Thaler (1999) found behavior in the opposite direction of Samuelson's colleague in two out of their three studies.

Using the one-parameter functional form for Barney, calibrations suggests that NBLLN goes much of the way, but *not* all of the way, in explaining why 25-57% of participants turned down the repeated bets in Benartzi and Thaler's studies. For Tommy, the coefficient of loss aversion required to explain this data is absurdly high, in excess of 32,000. For Barney with $\psi = 10$, the required loss aversion is approximately 15—many order of magnitudes closer to reality but still much larger than reasonable estimates of λ , such as 2.25.²⁶

In addition to predicting that people perceive too much risk in repeated betting, NBLLN also predicts that people perceive too much risk in the closely-related context of long-term investing. Indeed, analogously to the evidence on repeated betting, Benartzi & Thaler (1999) found that university employees vastly overestimated the probability that equities would lose money over a thirty-year horizon. Moreover, the employees stated a far greater willingness to invest in equities when they were explicitly shown the thirty-year returns. While there are many reasons why individuals may invest less in equities for retirement than recommended by standard finance models, we suspect that NBLLN is an important contributing factor. As such, just as in other settings researchers may underestimate risk aversion by ignoring overconfidence when inferring risk preferences from investment behavior, researchers might therefore exaggerate the risk aversion of investors by ignoring NBLLN.

²⁶Incorporating some of the biases missing from our model of NBLLN—that subjects still have excessively fat tails in their subjective sampling distribution helps reduce the required level of loss aversion even more—although still not to 2.25. Indeed, even if an individual exhibited the most extreme form of NBLLN *and* fat tails, putting equal weight on every possible outcome of the repeated gamble, the required level of loss aversion would be around 4 (because only 1/3 of the outcomes are losses). This exercise implies that some other bias, such as probability weighting, is also implicated in turning down the repeated bet. This may also explain why some subjects turn down the aggregate bet even when it is presented in histogram form.

4 The Multiple-Sample Model

In Section 2, we described how Barney processes a set of signals that he treats as one sample. In doing so, we deferred two questions that are crucial to answer before reaching conclusions about the implications of NBLLN in most economic settings. First, when and how does Barney separate out data into more than one sample, both as a function of how the data is presented to him, or how he might endogenously group information during his thinking and decision-making? Second, what does Barney believe about how he will process information in the future? In this section, we provide a framework for thinking about about these questions, and we propose possible answers. We review in Appendix B the scant and somewhat contradictory evidence we can find; since there is so little evidence about which assumptions are appropriate, our proposals are tentative. Nonetheless, we formulate plausible approaches in order to provide guidance for future experimental work, and to study the range of possible consequences of NBLLN in some dynamic decision-making environments.²⁷

For Tommy, it does not matter whether he treats 20 independent signals as one sample of 20, two samples of 10, or 20 samples of 1. And it is inherent to Tommy that he correctly predicts how he will treat future information. In contrast, and intrinsic to the very meaning of NBLLN, Barney's beliefs about the distribution of and inference from signals depend on how he divides them up into samples. The range of possibilities raises, in turn, the possibility that Barney predicts his future information-processing differently than how he will actually process information.

Several distinctions and definitions are usefully formulated to address these issues. We use the term *clump* to refer to how signals are objectively delivered to Barney by his environment. For example, when Barney asks a sequence of friends about their experience driving a Volvo, each friend's report arrives as a clump, but when Barney reads a summary of 10,000 individuals' experiences in *Consumer Reports*, the 10,000 signals arrive as a single clump. We refer to how Barney subjectively *processes* these clumps for the purposes of making forecasts and inferences as how he *groups* the signals. Barney forms beliefs regarding each group of signals as if a subjective rate β were drawn that applies only to that group, and hence the single-clump model from Section 2 can be applied to each group. In formulating ways that Barney might process information, we will distinguish two facets of how Barney groups data. The first is how he processes clumps into groups *retrospectively*—how he processes clumps he has already received. The second is how he processes clumps into groups *prospectively*. Prospective grouping determines his forecast about what data he

²⁷Moreover, the implications of other non-Bayesian models of judgment biases—such as base-rate neglect—similarly exhibit sensitivity to how data are framed. As such, the range of approaches we outline here may prove useful for studying those other biases.

will observe given his current beliefs *and* his forecast of what he will infer from that data after he observes it.

In each of the retrospective and prospective directions, we consider three ways that Barney might process clumps of signals. If Barney groups the signals the same way he receives them from the environment, we call him *acceptive*. Acceptive Barney would process *Consumer Reports* as a single sample of 10,000, and then each of his friends' reports as a separate sample. If Barney processes all of the clumps of signals he observes as a single, large group, we call him *pooling*. Pooling Barney would treat *Consumer Reports* data and his friends' stories as a single, larger sample. Finally, if Barney processes each signal individually (as a group of samples of size 1) no matter how they are clumped, we call him *atomizing*. Since Barney treats samples of size 1 identically to Tommy, if Barney is atomizing both prospectively and retrospectively, he is Tommy. If he does so in only one direction but not the other—examples of which we analyze in subsequent sections—he may be very different from Tommy.

To formalize the information-processing assumptions, consider a decision problem in which up to $1 \leq T \leq \infty$ binary signals will be realized in total. To greatly simplify notation and to sharpen the conceptual framework, we assume that how the environment or his own thinking leads him to group signals is independent of the realizations of those signals and of earlier decisions.²⁸ We define $T + 3$ partitions, each represented by a set of time periods, of the set of signals that fully characterize our predictions. We conceive of the first three of these partitions as embodying physical, informational, and perceptual assumptions rather than assumptions about Barney's statistical reasoning. The first partition we define by the set of dates at which the agent knows he must make decisions, $D \subseteq \{0, 1, \dots, T\}$. If $\tau \in D$, the agent knows that he made or will make a decision after observing τ signals but before observing $\tau + 1$ signals. The agent's payoff may depend on any or all of the decisions he makes, the signals that are realized, and the underlying state with which the signals are correlated. This opportunity-for-decisions partition, D , is of course specified in every economic model of decisionmaking.

The second partition is characterized by a set $C \subseteq \{0, 1, \dots, T - 1\}$ of dates where a new clump begins, such that if and only if $\tau \in C$, then signal τ is in a different clump than signal $\tau - 1$. We assume $0 \in C$. For example, if 11,000 signals arrives as a clump of 10,000 signals followed by a clump of 1,000 signals, then $T = 11,000$ and $C = \{0, 10, 001\}$. There is one obvious basic

²⁸This is a substantive restriction, and we know of examples where it seems unrealistic, but we do not know how relaxing it will improve insights. Importantly, we are not assuming that *whether* Barney gathers more information is independent of what he has learned; our examples in Section 5 revolve around exactly such a decision by Barney. In those examples, Barney faces a decision after each signal whether to pay to observe another signal. To accommodate those examples in the framework of this section, it can be assumed that whenever Barney chooses not to observe further signals, decisions that “occur” at those future signals do not affect his payoff.

restriction on clumping that we must make:

Clumping Assumption 0. $D \subseteq C$.

This assumption states that at any history where the agent makes a decision, subsequent signals arrive as a separate clump than previous signals. This restriction is inherent in the notion of clumps because the signals cannot have “arrived” together (at least in the relevant sense of the agent’s knowledge of their realizations) if the agent knows the realizations of only some of the signals. We treat this as a coherence assumption and always impose it.

Another coherence assumption on C is that Barney will not treat signals differently when he does not see the signals distinctly at all. If it is the case, unbeknownst to Barney, that 38 of the people in *Consumer Reports* statistics of 10,000 car owners have the last name Smith, we assume that Barney cannot (even if thusly motivated) treat Smiths as one sample and non-Smiths as another.²⁹ Let $I \subseteq \{0, 1, \dots, T - 1\}$ be a set of dates that defines the third partition, a partition of signals into equivalence classes of indistinguishable signals (so that if $\tau \in I$, then the signal at time τ is distinguishable from the signal at time $\tau - 1$). Clearly it must be the case that $D \subseteq I$, but we also impose a second restriction on clumping:

Clumping Assumption 1. $C \subseteq I$.

Although economic models of decisionmaking do not traditionally specify a clumping partition, our aspiration is have the clumping partition be an exogenous assumption that is not *per se* related to NBLLN, and ideally is pinned down by observable characteristics of a situation. In the interest of minimizing the number of assumptions that have to be specified anew for each new economic model (and hence limit degrees of freedom that might reduce the usefulness of the model), it would be especially attractive to pin down some rule for “clumping” that ties it to D or I . The two obvious candidates are:

Clumping Assumption 2(a). $C = D$.

Clumping Assumption 2(b). $C = I$.

While the three partitions characterized by D , C , and I reflect the physical and perceptual environment facing Barney, the remaining T partitions embed Barney’s NBLLN psychology of how he separates out data. To capture the important possibility that Barney’s grouping might differ

²⁹Conceivably, Barney could decide to label an indistinguishable group any way he wants, such as ordering them from 1 to 10,000, and then either perceptually or psychologically distinguish the signals based on this labeling. It seems a safe assumption that he will not do so.

from different time perspectives, we assume that the agent may process the signals differently at different dates. At each $t = 0, 1, \dots, T$, there is a set $P_t \subseteq \{0, 1, \dots, T - 1\}$ of dates where a new group begins. By letting each P_t contain elements both less than t and greater than t , each partition captures both retrospective and prospective grouping of signals at that point in time. For all $\tau \leq t$, if $\tau \in P_t$, then after having observed t signals, the agent processes signal τ as being in a different group than signal $\tau - 1$; while if $\tau \notin P_t$, then after having observed t signals, the agent processes signal τ as being in the same group as signal $\tau - 1$. And for all $\tau > t$, $\tau \in P_t$ means Barney *anticipates* separating out signal τ from signal $\tau - 1$. Whether he actually does so after observing signal τ is determined by $\{P_\tau, P_{\tau+1}, \dots\}$. We assume that $0 \in P_t$ for all t .

We impose one restriction on these sets that we believe is necessary for modeling coherence:

Processing Assumption 0. For any $t = 0, 1, \dots, T$, if $\tau \in D$ and $\tau + 1 \geq t$, then $\tau + 1 \in P_t$.

This assumption states that at any date where the agent makes a decision, he processes signals before and after that date as being in separate groups—and that before that date, he knows he will do so. We consider this to be a modeling coherence assumption because it ensures that Barney’s NBLLN from the single-clump model in Section 2 generalizes to every decision node in the multiple-clump model. To see this, suppose Barney knows that $\theta = .5$, reads a summary of 10,000 individuals’ experiences in *Consumer Reports*, and then must make a prediction about the next 1,000 signals he will observe. If, in violation of Processing Assumption 0, he were to process all 11,000 signals together as a single group after already observing the first 10,000 signals, then he would believe that the same subjective rate β applies to all 11,000 signals. Using the first 10,000 signals, he would update his beliefs about β from $f_{\beta|\Theta}^\psi(\beta|\theta = .5)$ to a density that puts almost all the probability mass on the observed proportion of a signals, say 50%. Since the next clump is grouped with the earlier clump, his subjective sampling distribution for the next clump will put negligible weight on a proportion of a signals outside a neighborhood of 50%. In his predictions about future signals, Barney would no longer exhibit NBLLN. In contrast, Processing Assumption 0 requires that Barney forms beliefs as if a new β is drawn from $f_{\beta|\Theta}^\psi(\beta|\theta)$ before the next 1,000 signals, so his subjective sampling distribution is exactly as in the single-clump model for $N = 1000$. Precisely because it rules out learning about the subjective rate, Processing Assumption 0 distinguishes our model of NBLLN from the generalization of the model from Section 2 that one would employ if one interpreted it as a fully-rational model with uncertainty about the parameter β .

We can now formalize various ways that Barney might form beliefs retrospectively and prospectively. Because D is the set of nodes where Barney’s beliefs are payoff-relevant, these definitions focus the assumptions on the P_t ’s where $t \in D$:

- **Retrospective-Pooling:** If $t \in D$ and $0 < \tau \leq t$, then $\tau \notin P_t$.
- **Retrospective-Acceptive:** If $t \in D$ and $0 < \tau \leq t$, then $\tau \in P_t \Leftrightarrow \tau \in C$.
- **Retrospective-Atomizing:** If $t \in D$ and $0 < \tau \leq t$, then $\tau \in P_t$.
- **Prospective-Acceptive:** If $t \in D$ and $\tau > t$, then $\tau \in P_t \Leftrightarrow \tau \in C$.
- **Prospective-Atomizing:** If $t \in D$ and $\tau > t$, then $\tau \in P_t$.

We omit defining “prospective-pooling” Barney because it is ruled out by Processing Assumption 0 in any decision problem with more than one decision node; at an earlier decision node, Barney cannot expect to pool together future signals that come before and after a future decision node. (Barney’s predictions regarding a single clump and a single decision node, as in Section 2, are of course covered by the prospective-acceptive case.)

As an implication of all the other ways he is rational, Tommy always processes information the way that he expects to process information. We call this property “processing-consistency”: An agent is **processing-consistent** if $P_t = P_{t'}$ for all $t, t' \in D$. Despite his irrationality, Barney shares this property if his retrospective and prospective thought processes coincide. In particular, if Barney is retrospective-acceptive and prospective-acceptive, then he is processing-consistent and accurately forecasts what his own future beliefs will be after he observes a sequence of signals.³⁰ In contrast, if Barney is retrospective-pooling and prospective-acceptive, then he is not processing-consistent. As a result, he may behave in a time-inconsistent way, e.g., expecting to learn a lot from purchasing a large number of signals, but remaining uncertain after observing the signals and therefore preferring to purchase yet more signals. This time-inconsistency will play a role in some of the applications we study.

Having formalized the processing assumptions we will focus on, we now discuss intuitions not captured by the formalism. When thinking about what processing assumption is psychologically plausible, we try to be guided by the characteristics of the data-generating process. For example, friends’ vivid descriptions of their own personal experiences are easily distinguished from each other and therefore are likely to be processed as separate groups. Indeed, here we note questions and clarifications that our framework can add to how some phenomena have been conceptualized in psychological research. A tradition in psychology emphasizes that people may overweight “vivid” evidence in reaching their judgments; a friend’s colorful description of the horrors that ensued when her car broke down while trying to pick up her child from school may weigh more heavily in our

³⁰Note that acceptiveness is the sole processing-consistent case we focus on because we rule out prospective pooling, and a processing-consistent atomizer is named Tommy.

judgment of the brand of car she drives than summary statistics based on large samples of data like those sometimes published in *Consumer Reports* (e.g., Nisbett & Ross, 1980). Our model predicts a version of this if vivid signals are isolated from less vivid ones. But the *comparative* over-use of vivid evidence may not be over-use of that evidence relative to its proper use, but an indication of under-use of other evidence. The observation that you are using your friend’s car experience a lot more than other, similar people’s experiences that show up in the data could be, in light of NLLN, partly or even fully attributable to the under-use of statistical data.³¹ Moreover, even when it seems clear that there is genuine over-use of evidence due to vividness, NLLN is still clearly an “enabling bias.” The simple logic that when a person has large data sets she ought to reach a determinative conclusion means that any sort of inference (rational or irrational) from small, additional samples simply should not matter. Counting a friend’s car experience a hundredfold still should not influence beliefs by much.

In other cases, it is harder to see whether and when Barney will group signals together, and (importantly in our applications) whether he is more likely to pool signals prospectively or retrospectively. Maybe he pools information prospectively (in violation of Processing Assumption 0) but atomizes it retrospectively: before he talks to the 10 friends (say) he plans to talk to, he may not attend to the time separation of the information and not realize that he will update his beliefs story-by-story as he goes along. But then he may retrospectively atomize, distinguishing colorful details of his friends’ stories. Alternatively, maybe he instead processes information more coarsely retrospectively than he anticipates, as we assume in an example in Section 5: if he plans to keep talking to friends one by one until he feels confident, he might think ahead with attention to each separate signal, focusing on how he will update from current beliefs after his next conversation. But then in retrospect, after he has talked to his next friend, he may quite naturally treat that friend’s information symmetrically with all the previous conversations and take stock of his current information by thinking together about all the advice he has received.

We hypothesize that many situations are like the last example, and that the psychology of grouping signals leads Barney to believe he will process data more finely in the future than he actually will do retrospectively. For example, while Barney might recognize that he receives a new signal about the average precipitation of San Francisco every day, and atomizes the signals prospectively, when he attempts to infer whether San Francisco is “rainy” or “sunny,” he treats all

³¹This hypothesis that NLLN may help explain why some signals may receive attention in ways that seem surprising may shed a favorable light on some important research lines. For instance, the influence on beliefs of friends’ information and experiences, learning by one’s own experience, and learning in networks and herds, might seem puzzling in light of the amount of other information an agent may have available. But if information from the rest of society is treated as just one sample, the information provided by members of a the agent’s more immediate social network may still loom large.

previous outcomes in a coarser fashion, possibly grouping days by weeks, months, or all together.

We have outlined above a few possibilities about how Barney might process clumps of signals, but there are other reasonable possibilities, some of which violate our assumptions. In particular, we emphasize that acceptive Barney is only one plausible type of Barney that is between pooling and atomizing. Barney could process information in a different fashion than presented by his environment, grouping it instead according to the perceived similarity of the information source. After observing 10,000 datapoints from *Consumer Reports* followed by 10 friends' reports obtained sequentially, for instance, Barney may retrospectively process the information as a group of 10,000 followed by a group of 10.

Besides lack of evidence, an additional and major reason the hypotheses in this section are tentative is that we have completely sidestepped the issue of when and how Barney might “think through” his beliefs more fully. Even if Barney is not processing-consistent, given the assumptions above, our *model* of Barney’s beliefs is internally consistent. Barney’s *beliefs* themselves, however, are not internally consistent, and this raises additional conceptual and practical issues in applying a model of NBLLN.³² For example, a teacher could elicit Barney’s belief about the likelihood that a first signal will be a , the likelihood that a second signal will be a conditional on the first signal being a , and the likelihood that a sequence of two signals will be aa ; then the teacher could point out that the product of the first two does not equal the second. In fact, even our coherence assumptions above could fail depending on the questions a teacher asked Barney. For example, suppose Barney expects to observe 10,000 signals from *Consumer Reports*, then make some payoff-relevant decision, and then observe another 1,000 signals. If a teacher asks Barney to forecast all 11,000 signals, then Barney would presumably do so according to the single-clump model with $N = 11,000$ —in violation of Processing Assumption 0. Even in the absence of a “teacher,” Barney might ask himself such questions.³³ While we flag these issues, and think they are natural subjects of future research,

³²The internal inconsistency we highlight here does not arise in “false-model Bayesian” models of biased beliefs such as Barberis, Shleifer & Vishny (1998), Rabin (2002), and Rabin & Vayanos (2010). In these models, biases are formulated as agents holding the wrong theory as to the statistical structure of the world, but as being fully Bayesian in their interpretation of data within that structure. So long as all events that are possible in the true world are also possible in the agents’ imagined world, no internal inconsistency can arise. (Though even in these models, it may be very likely that the agent will observe a sequence of signals that he perceives to be very unlikely.) Processing-consistent variants of Barney likewise reduce to a “false-model Bayesian” theory. The processing-*inconsistent* variants of NBLLN, however, assume an intrinsically non-Bayesian thought process. The modeling challenges associated with internally inconsistent beliefs are not specific to NBLLN, and will arise in any model of belief formation that is fundamentally non-Bayesian.

³³Importantly, a decision Barney faces might itself naturally cause Barney to ask himself such questions. For example, imagine Barney is making a decision whose payoff depends on whether the state is A or B . He could purchase one signal, and then decide whether to purchase a second signal, or he could purchase two signals all at once at a discount. When deciding what to do, it seems natural that Barney would ask himself what he would conclude after observing each of the three possible outcomes: he observes 1 signal, 1 signal followed by 1 signal, and 2 signals together. Having explicitly asked himself about these possibilities, it seems odd that Barney would—as assumed if Barney is prospective-acceptive—expect to conclude less

we proceed in subsequent sections with the assumption that Barney does not think through the inconsistencies in his own beliefs.

5 Application: Search and Experimentation

This section explores the economic implications of NLLN inference, applying variants of our model both to classical information acquisition and to experimentation. In the applications here and in Section 6, we use a standard framework, along the lines of many examples above. An agent is uncertain about which of two possible states of the world, $\omega \in \{A, B\}$, is true. State A has prior probability $0 < f_{\Theta}(\theta_A) < 1$, and state B has prior probability $f_{\Theta}(\theta_B) = 1 - f_{\Theta}(\theta_A)$. In state A , the probability of an a -signal is θ_A , and the probability of a b -signal is $1 - \theta_A$. In state B , the probability of an a -signal is $\theta_B < \theta_A$, and the probability of a b -signal is $1 - \theta_B$. Depending on the particular application, the agent can take actions, or observe outcomes and signals that can inform him about the state of the world. Here and in Section 6, we assume that Barney’s subjective rate distribution has the functional form (3) from Section 2, although we believe the intuitions are more general. The applications make use of the varying assumptions defined in Section 4 about how Barney groups signals retrospectively and prospectively. While the analysis of betting and investing focused only on the implications of Barney’s subjective sampling distribution, Barney’s behavior in the current setting depends both on his subjective sampling distribution—what signals he expects to observe—and his inferences—what he expects to believe and actually believes after observing the signals.

An interesting implication of NLLN emerges across a range of applications: it makes a bigger difference to his eventual beliefs for Barney than for Tommy whether he happens to observe strong evidence of the true state early or late in his learning process. The basic logic of NLLN implies that Barney finds observing strong, early evidence, such as the group of signals aa , more persuasive that A is the true state than a group such as $ababaa$, even though these two groups are, objectively, similarly strong signals about the state—indeed, exactly equally strong when $1 - \theta_A = \theta_B$. As a result, if Barney observes strong evidence early on, he may stop trying to learn about the state after only a few signals, while if he observes ambiguous data early on, he may continue trying to learn even after many signals.

Consider a setting of sequential information acquisition. Barney is trying to decide what make of car to buy, a Volvo or a Lada.³⁴ He could possibly incur costs to acquire signals about the relative quality the cars, e.g., by taking time to ask friends which car is better. Conditional on the

from the 2 signals together than the 2 signals individually.

³⁴A Lada is a type of car. So is a Volvo.

responses that he receives, he can decide to ask more friends, or to stop and choose a car to buy.

Each period $t = 1, 2, \dots$, the agent can choose to purchase a single signal at cost $c > 0$ or take an action $\mu \in \{\mu_A, \mu_B\}$.³⁵ If the agent takes an action, he gets payoff $u(\mu, \omega)$, which equals 1 if the action matches the state ω and 0 otherwise, and the agent faces no further decisions. If the agent decides to purchase an additional signal, he sees the realization of the signal, and he proceeds to the next period. The agent will live forever and seeks to maximize the expected action payoff minus expected signal-purchase costs; if the agent purchases κ signals and then takes action μ , his utility is $u(\mu, \omega) - \kappa c$. Note that we assume no discounting, so that the only reason an agent would stop acquiring information before being absolutely certain is the cost c of obtaining an additional signal.

For Tommy, the characterization of optimal behavior is well-known (e.g., Wald, 1947). Each time Tommy purchases a signal, he updates his posterior beliefs. His optimal behavior is characterized by two probabilities, ν_l and ν_h , with $0 < \nu_l < \nu_h < 1$. If and only if the posterior probability of state A exceeds ν_h , he stops and takes action μ_A ; if and only if it goes below ν_l , he stops and takes action μ_B . Tommy continues to purchase signals as long as his posterior beliefs remains between ν_l and ν_h . But because his posterior ratio is a martingale process, Tommy will eventually feel strongly enough to take an action almost surely. Importantly, as $c \rightarrow 0$, $\nu_l \rightarrow 0$ and $\nu_h \rightarrow 1$.

Barney's behavior will depend on how he groups signals. Because the signals arrive one at a time, Barney expects to group the signals as samples of size 1, regardless of whether he is prospective-atomizing or prospective-acceptive. Since Barney *expects* to behave exactly like Tommy, his policy is the same as Tommy's, with the *same* thresholds ν_l and ν_h determining when he will stop and take an action. However, Barney's posterior belief may follow a different process than Tommy's, depending on how he groups signals retrospectively. If Barney is retrospective-atomizing or retrospective-acceptive, his beliefs and behavior will be identical to Tommy's.

If he is retrospective-pooling, however, Barney's behavior can differ qualitatively from Tommy's. In this case, the impact of an additional signal on his posterior beliefs is smaller than for Tommy. Hence in expectation, Barney will purchase weakly more signals than Tommy. Moreover, the marginal impact of an additional signal on Barney's posterior beliefs will approach zero as his sample of observed signals grows. This is because Barney's inference becomes more and more driven by the proportion of a signals, which is less affected by an additional signal in a larger sample. However, Barney *believes* that an additional signal will have the same impact, regardless

³⁵In all our applications, we restrict Barney to purchasing information a single realization at a time. If we allowed Barney to choose how many signals he could purchase each period, Barney would have to think about what he would learn from purchasing two signals sequentially in order to compare it to purchasing two signals simultaneously. Modeling this thought process raises challenges—the same as those mentioned in footnote 33—that we sidestep in this paper.

of the sample size he has already observed. As a result, Barney can become stuck in a *learning trap*, in which he purchases signals forever, but they will never change his confidence in the state of the world enough for him to stop. Proposition 7 summarizes this discussion:

Proposition 7. *Fix payoffs $u(\mu, \omega)$, rates $\theta_A > \theta_B$, and priors $f_{\Theta}(\theta_A)$ and $f_{\Theta}(\theta_B)$. Suppose Barney is prospective-atomizing or prospective-acceptive.*

1. *If Barney is retrospective-atomizing or retrospective-acceptive, then for any $c > 0$, Barney's behavior exactly coincides with Tommy's, and he will choose an action after observing a finite number of signals almost surely.*
2. *If Barney is retrospective-pooling, then for all $\bar{p} < 1$, there exists $\bar{c} > 0$ such that for all $c \leq \bar{c}$, Barney purchases an infinite number of signals with probability $p > \bar{p}$.*
3. *Suppose Barney is retrospective-pooling, $\theta_A = 1 - \theta_B$, and Barney is willing to purchase an additional signal when his posterior probability of state A equal to q . (a) If Barney's posterior probability of state A is q after observing N signals and again q after observing $N' > N$ signals, then the probability that Barney will purchase an infinite number of signals is weakly higher after he has observed the N' signals than it was after the N signals. (b) Furthermore, there exists a \hat{N} such that for all $N' > \hat{N}$, the probability that Barney will purchase an infinite number of signals is strictly higher after he has observed the N' signals than after the N signals.*

Since the marginal effect of a signal is shrinking toward zero, Barney's posterior probability of state A has a positive probability of always remaining between ν_l and ν_h . If Barney ends up in such a learning trap, then his welfare is unboundedly negative. In fact, because a small c tempts him to wait longer, the probability of this unboundedly negative welfare becomes arbitrarily close to 1 as the signal cost c becomes arbitrarily small. For a given c , the last part of the proposition states that Barney is more likely to get caught in a learning trap the more signals he has already observed, holding constant his posterior belief. For example, if the rates θ_A and θ_B are symmetric, then an equal number of a and b signals does not change Barney's beliefs; so Barney is more likely to end up purchasing an infinite number of signals after having observed $abab$ than he was before he observed any signals.³⁶

³⁶We have found several experiments that set up a dynamic information-purchase setting with a payoff structure similar to the model in the text and that compare subjects' behavior with a Bayesian benchmark, which is calculated assuming expected-value maximization. Tversky & Edwards (1966), Pitz (1968), Wendt (1969), and Hershman & Levine (1970) found that subjects purchased too much information. In contrast,

We conclude this section by considering the implications of NBLLN in the related situation where an agent learns through experimentation. Rather than purchasing signals about the quality of the car, Barney could instead take them for test drives, perhaps by renting them. Here, instead of an explicit cost, the cost of information acquisition is the cost of waiting to purchase the correct car. If the Volvo is the better car, then Barney is losing out every day he drives the Lada. For the same reasons that retrospective-pooling Barney could get caught in a learning trap in a dynamic information-acquisition setting, he could end up taking a suboptimal action forever in an experimentation setting.

Formally, in each of an infinite number of periods, $t = 1, 2, \dots$, the agent takes an action $\mu_t \in \{\mu^1, \mu^2, \dots, \mu^J\}$ where $J > 1$. After taking an action, the agent receives the (possibly unobserved) payoff $u(\mu_t, \omega)$ and a signal that takes on a value of a or b . An a -signal is generated with probability $0 < \theta_{\mu_t, \omega} < 1$ that is partially revealing of the state (that is, $\theta_{\mu^j, B} < \theta_{\mu^j, A}$ for all μ^j). Here, we assume the agent discounts future utility flows at rate $0 < \delta < 1$. The problem is interesting when the agent faces a tradeoff between taking the action that maximizes his expected payoff subject to his current beliefs—the “myopically best action”—and others which, although giving a low expected payoff, can help reveal the true state of the world, which enables him to choose a better action thereafter.

In this environment, Tommy will eventually stop experimenting, and his behavior will settle down to a single action. To be precise, Aghion, Bolton, Harris & Julien (1991) showed that almost surely at some finite T , Tommy will take the action that gives the highest per-period payoff conditional on his current belief for all periods $t \geq T$.

Just as with information-acquisition, because he observes one signal at a time, Barney will expect to behave like Tommy regardless of whether he is prospective-atomizing or prospective-acceptive. Retrospective-atomizing and retrospective-acceptive Barney will behave identically to Tommy. However, retrospective-pooling Barney may continue to experiment forever with positive probability, with the same possibilities and intuitions as in Proposition 7 above.

There are examples of experimentation environments where there is positive probability that Tommy never learns the true state, e.g., when the myopically best action provides no information about the true state. However, when there are a finite number of possible actions, and each provides some information about the true state, Tommy will eventually discover the true state and take the

Fried & Peterson (1969) and Pitz & Barrett (1969) found that subjects purchased too little information. Moreover, Pitz & Barrett found that when the already-observed sample size was larger, holding constant the objective strength of evidence, subjects bought *fewer* additional signals. Also contrary to our model’s prediction, Sanders & Ter Linden (1967) Studies 1-3 found that, when the already-observed sample size was larger, subjects stopped acquiring information at a point where the objective evidence was weaker. In Sanders & Ter Linden’s experiments, however, the signals arrived at a rate of 2, 5, or 10 signals per second, which is so fast that the nature of the inference task is likely quite different than in other studies.

best action. By contrast, because Barney’s learning is limited even when he observes an infinite amount of data, he may never learn enough to take the correct action.

Barney’s failure due to NBLLN to learn in an experimentation environment may play an important role in allowing people to maintain misconceptions about themselves despite extensive feedback. For example, naivety about present-biased preferences—a false belief that, even though one weights current utility flows disproportionately relative to future utility flows, one will not do so in the future—is now widely considered to have important implications in a variety of economic settings, differing dramatically from sophisticated present-bias. It is natural to wonder how it can be that people do not eventually learn their true discount function by observing their own behavior.³⁷ The variant of NBLLN that assumes retrospective-pooling implies that people may not become certain of their own dispositions even after an infinite amount of self-observation. NBLLN may similarly be implicated in the persistence of other biases, such as overoptimism about one’s own abilities. NBLLN may also enable the otherwise-puzzling lack of knowledge of one’s own preferences that is presupposed in any self-signaling model.

6 Application: Observational Learning

Besides learning from his own experience or gathered information, an agent could also choose which car to buy by observing which car his neighbors have bought. Our final application concerns issues in modeling and potential implications of NBLLN in observational learning. Once again, we assume that there are two states, $\omega \in \{A, B\}$, with corresponding rates for a signals, $\theta_A > \theta_B$, and actions $\mu_t \in \{a, b\}$ (here we use the same notation for actions as for signals). The action pays off 1 if it matches the state and 0 otherwise. There is an infinite sequence of agents, $t \in 1, 2, \dots$, who each observe a single signal, choose an action in order, and receive a payoff. Agents have common prior beliefs $f_\Theta(\theta_A) = 1 - f_\Theta(\theta_B)$. In addition to their private signal, agents observe the action taken by each previous agent, and the order in which those actions are taken.

Much of the theoretical literature on social observational learning has focused on two closely-related phenomena, “information cascades” and “herds.”

Definition 1. *A cascade is said to occur when, following some period, each agent’s choice of action does not depend on the agents’ private signals.*

³⁷Ali (2011) formalizes this intuition in a planner-doer model and concludes that a Bayesian-rational agent who succumbs to temptations would eventually learn that he has a self-control problem. In analysis of this environment that we do not include here, we find that (a) Barney may forever remain uncertain about his self-control problem, despite succumbing to temptation every period; and (b) the probability of this persistent uncertainty—and hence optimism about his ability to resist temptation—is increasing in Barney’s ex-ante optimism.

Definition 2. *A herd is said to occur when, following some period, all agents choose the same action. It is a good herd if the common action is the true state; a bad herd if the common action is not the true state.*

The characterization of Bayesian Nash equilibria in this simple setting is well-known (e.g., Banerjee, 1992; Bikhchandani, Hirshleifer & Welch, 1992; and Smith & Sørensen, 2000). When the agents are all Tommy, in all equilibria both a cascade and a herd occur with probability 1. To see the intuition for why, suppose the common prior puts equal weight on each state, and $1 - \theta_A = \theta_B$. In that case, a cascade will occur as soon as the total number of a -actions taken exceeds or falls behind the total number of b -actions by 2. For example, if the actions from the first 8 agents are, in order, $abababaa$, then the 9th agent will believe the state is more likely to be A even if his private signal was b . He will play a regardless of his own signal. The 10th agent understands this logic, so if he observes $abababaaa$; he will likewise ignore his own signal and also understand not to infer anything from the 9th action. Similarly for the 11th agent, and so on. The cascade occurs in period 9, with all agents $t = 9, 10, \dots$ “herding” on action a . A good herd is of course more likely than a bad herd because early actions are indicative of the true state, but the herd will be on the wrong action with positive probability.

Although no agent directly observes more than one signal, NBLLN will influence herding under the assumption that agents apply their non-belief to the signals they infer others are getting. Intuitively, if Barney groups together previous agents’ actions, then NBLLN implies that he will infer less from their actions than he should. Therefore, he will more often rely on his own signal, making it slower for a herd to form. Moreover, since Barney needs to see more agents following their own signal before ignoring signals than does Tommy, when a herd *does* occur, it is more likely to be on the correct action. Finally, because Barney’s learning is limited even in an infinite sample, a qualitatively different kind of behavior is possible. Because Barney may vastly under-infer from a large number of actions, it may be possible that if agents do not herd quickly enough, they will instead form what we call an “eddy”:

Definition 3. *An eddy is said to occur when, following some period, every agent chooses the action corresponding to his own signal.*³⁸

Because agents do not care about future agents’ actions, the way Barney groups signals prospec-

³⁸Smith & Sørensen (2000) show that when agents are all Tommy, but have ordinally different preferences, an eddy can occur. However, in their model, this results from what they describe as “confounded learning”: agents’ actions, although dependent on their private signal, are uninformative about the state of the world because an action’s informational meaning is different depending on the (unobservable) preference of the agent. In contrast, we find that when agents are Barney, eddies occur even when actions are informative. Barney simply never learns enough from this information.

tively is irrelevant for his behavior. For the purposes of this section, we assume that Barney groups all previous agents' actions together and treats his own private signal as separate. This assumption corresponds to being "retrospective-acceptive" if Barney observes the previous actions simultaneously, as well as the order that the actions were taken in, and then observes his own signal. We think this grouping assumption would, in fact, be natural even if agents observed their own signals at the same time as observing actions because other agents' actions may feel like qualitatively different information than the agent's own signal.³⁹

Deriving the full implications of NBLLN in this setting requires an extension of the framework introduced in Section 4 to confront an issue that will be important in many economic applications of NBLLN, as it is for other models of cognitive biases. Namely, what is Barney's theory as to how *other* agents draw inferences? While a complete exploration of Barney's beliefs about others' beliefs is beyond the scope of this paper, the herding environment can illustrate some of the assumptions we think may be reasonable. In the analysis here, we compare two different assumptions about Barney's beliefs about other agents' inferences, both of which seem plausible. We say that Barney is *strategically sophisticated* if, when he draws inferences from other agents' behavior, he assumes that they group signals the way they in fact do group signals. That is, even though the 3rd agent groups together the actions of the 1st and 2nd agents, he understands that the 2nd agent does not group his own signal with the 1st agent's action.⁴⁰ Alternatively, we say that Barney is *strategically naive* if, when he draws inferences from other agents' behavior, he assumes that they group signals the way he himself does. In that case, the 3rd agent incorrectly believes that the 2nd agent groups his own signal with the 1st agent's action.

Although we believe that strategic naivety is a plausible assumption, in the current framework it generates significant problems. It is possible that strategically naive Barney can observe sequences of actions that he cannot rationalize.⁴¹ For example, consider a set of strategically naive Barneys, where $\theta_A = .51 = 1 - \theta_B$, and $\psi = .10$. The first 100 Barneys receive alternating a and b signals in order (and so take alternating a and b actions). The 101st and 102nd agents both receive a b signal, and so both take a b action. From the perspective of the 104th agent a herd should start on action b with the 103rd agent. However, if the 103rd agent receives an a signal, he will still take an

³⁹The assumption about grouping is again important. If instead, for example, we assumed Barney pooled his own signal with all other inferred signals, then the standard results in the literature carry through: a cascade will always occur, and therefore a herd will happen, sometimes on the incorrect action.

⁴⁰Although strategic sophistication implies that each agent understands that other agents are using a different partition than themselves, they can support such beliefs by thinking that other agents are wrong.

⁴¹In a closely related model that we have analyzed—where agents only observe the total number of each type of action, but not the order in which they were taken—this problem never arises. The results presented here for sophisticated agents apply in that setting for both strategically sophisticated and strategically naive Barney. In addition, if all agents are strategically naive Barney, then there are parameter values such that there exist finite histories (reached with positive probability) after which an eddy occurs with probability 1.

a action, as his own signal outweighs the informativeness of the pooled actions of previous agents. In this case, the 104th agent will not be able to rationalize the observed sequence of outcomes. We sidestep this problem by focusing solely on sophistication.

We begin the formal analysis by characterizing when Barney herds. Although models of observational learning admit multiple equilibria, we will follow the literature in assuming that when an agent is indifferent between actions, which occurs when his own signal balances the information from previous actions (and so the agent is indifferent between actions), the agent follows his own signal.⁴² Proposition 8 states that if the prior puts sufficient weight on one of the states, then in equilibrium there is positive probability of a herd on that state. The reason (but not the precise thresholds) is exactly the same for Barney as for Tommy: if early agents receive enough signals in favor of the state, say A , and take action a , then eventually it makes sense for Barney (and all subsequent Barneys) to take action a regardless of his signal. However, if the prior in favor of A is too low, then it is impossible to herd on action a because the game will begin with a herd on b .

Proposition 8. *Suppose Barney is strategically sophisticated. Fix rates $\theta_A > \theta_B$ and Barney's parameter $0 < \psi < \infty$. There are bounds $0 < \underline{p}_A < \frac{1}{2}$ and $\frac{1}{2} < \bar{p}_A < 1$ such that if the common prior $f_\Theta(\theta_A) \geq \underline{p}_A$, there is positive probability of herding on action a ; and if $f_\Theta(\theta_A) \leq \bar{p}_A$, there is positive probability of herding on action b .*

The proposition implies that, like for Tommy, there exists a range of priors ($\bar{p}_A \leq f_\Theta(\theta_A) \leq \underline{p}_A$) such that in equilibrium there is positive probability of a herd on either state.

We next characterize when an eddy might occur. An eddy can occur for Barney but *not* Tommy. If a herd has not occurred after many actions, then the early signals must have provided mixed evidence. For Tommy, as long as $\theta_A = 1 - \theta_B$, a herd starts after one action is played twice in a row, whether this occurs after two actions or eight actions. In contrast, Barney would draw a weaker inference following *abababaa* than *aa*. Hence, if the early evidence is mixed, then the probability of a herd falls because even strong evidence thereafter will not affect Barney's beliefs by as much. In fact, the probability of a herd starting vanishes as the sample size of mixed actions gets large.

For an eddy to occur, it must be the case that beliefs do not become too extreme initially, due to either extreme priors or lopsided initial signals. But it is also necessary that the prior put low enough weight on the true state that even in a very large sample, where the Law of Large Numbers ensures that the proportion of a signals will exactly match the true rate, Barney's inference from other agents' actions is weak enough that he still follows his own signal. For any rates $\theta_A > \theta_B$ and

⁴²Other equilibria, where the agent with some probability takes the action opposite his signal, will reduce the amount of publicly available information. In the extreme case, where an agent always takes an action opposite his signal when indifferent, a herd always starts in the first period.

Barneyness parameter $0 < \psi < \infty$, given the true state, there exists a range of prior beliefs such that an eddy occurs with positive probability.⁴³ Note that this range of prior beliefs is different depending on the true state, and may or may not overlap for the two states.

If the two states are close enough to being equally likely, and if the rates θ_A and θ_B are close enough together, however, then the range of prior beliefs that enables an eddy in state A definitely overlaps with the range in state B , and there is positive probability of an eddy in both states. Under those conditions, Barney’s posterior belief will remain weak after observing a large number of previous players’ actions, regardless of whether the proportion of a actions is θ_A or θ_B .

Proposition 9. *Suppose Barney is strategically sophisticated. There exist $\delta > 0$ and $\gamma > 0$ such that if $|f_{\Theta}(\theta_A) - f_{\Theta}(\theta_B)| < \delta$ and $|\theta_A - \theta_B| < \gamma$, then there is positive probability of an eddy regardless of which state is true.*

In most cases when an eddy occurs, there is positive (but shrinking) probability along the equilibrium path that the potential eddy will end and be superceded by a herd. The next proposition formalizes this: the probability of a herd vanishes as the sample size gets large because the actions of later agents have negligible impact on the proportion of a actions and hence negligible impact on Barney’s beliefs. Furthermore, as the sample size gets large, the relative likelihood of a bad herd to a good herd vanishes.

Proposition 10. *Suppose Barney is strategically sophisticated.*

1. *For any $\varepsilon > 0$, there exists $T_{\varepsilon} > 2$ such that if a herd has not occurred by period T_{ε} , the probability of herding from then on is less than ε .*
2. *Furthermore, suppose $1 - \theta_A = \theta_B$ and the prior $f_{\Theta}(\theta_A) \geq .5$ then for any $\nu > 0$ there exists $T_{\nu} > 2$ such that if a herd has not occurred by period T_{ν} , the likelihood ratio of a bad herd to a good herd is less than ν .*

Although in an eddy it will never happen that all agents will take the incorrect action (as in a bad herd), it is the case that eddies are on average less efficient than herds. To see this, again consider the case when signals are symmetric. The ratio of correct actions to incorrect actions in a herd is simply the ratio of good to bad herds. This is equivalent to the ratio of the probability of two

⁴³The intuition that there are parameter values where an eddy occurs with positive probability is robust to alternative models of strategic behavior, such as Eyster & Rabin’s (2010) model of inferential naivety. This is because the situations where Barney’s behavior causes eddies—sequences of signals where everyone’s action reveals their signals—are those situations where inferential naivety does not affect agents’ decisions.

signals that match the state (which we suppose is A) to the probability of two signals that do not match the state, $\left(\frac{\theta_A}{1-\theta_A}\right)^2$. In an eddy the ratio of correct actions to incorrect actions is simply the ratio of the probability of a correct signal to the probability of an incorrect signal, $\left(\frac{\theta_A}{1-\theta_A}\right)$. Since $\frac{\theta_A}{1-\theta_A} > 1$, the average payoff is higher in a herd than in an eddy.

7 Conclusion

The logic of NBLLN unambiguously predicts that people will extract far too little information from large samples, but there are strands of literature both within psychology and within economics on “over-confidence” in beliefs. In our reading of the evidence in Appendix B, it is clear that people tend to be over-confident in some environments and under-confident in other environments. Rather than viewing over-confidence and under-confidence as fundamental biases in themselves, we view both as outcomes to be explained as a function of the information a person is confronted with. The model of NBLLN highlights a feature of the decisionmaking environment—namely, sample size—that affects the degree to which an agent will draw too weak an inference from evidence. In Appendix A we combine NBLLN with the Law of Small Numbers (LSN) bias, which generates a bias toward over-confidence in inferences, and the overall pattern we predict is over-inference in small samples and under-inference in large samples.

There are conceptualizations of the tendency to under-infer from large samples that differ from that embedded in our model. 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. For example, in the case of under-inference, Winkler & Murphy (1973) posit that people treat independent signals as if they were positively correlated because their real-world experience is with positively correlated signals. Such positive correlation would generate excessively-dispersed subjective sampling distributions and under-inference but *not* NBLLN: the Law of Large Numbers applies to positively correlated signals, so an agent who misperceived independent signals as positively correlated would still believe that proportions converge to the population rate. Moreover, while ecological mismatch arguments often have merit, we think the argument is unappealing in this context because the bias we call NBLLN is evident in examples with which subjects have a great deal of real-world experience, such as coin-flipping.⁴⁴

Many have proposed conceptualizing NBLLN as one consequence of the “representativeness

⁴⁴We also note that in the case of the Law of Small Numbers, the opposite ecological-mismatch hypothesis is often proposed: that people ordinarily deal with *negatively*-autocorrelated signals. Typical real-world processes would have to have a fairly complicated form involving short-run negative autocorrelation and long-run positive autocorrelation to rationalize both the Law of Small Numbers and NBLLN.

heuristic,” according to which people draw inferences based on the degree of similarity between features of a sample and features of a population from which the sample might have been drawn. Indeed, Kahneman & Tversky (1972) present evidence for N BLLN in precisely this context. Although N BLLN certainly seems consistent with representativeness, it is not clear how the logic of representativeness predicts the prototypical case of N BLLN: e.g., an agent who observes 6 million heads and 4 millions tails continues to put non-trivial probability on the coin being fair. Representativeness could explain this kind of observation if it is interpreted as inferences based on proportions, combined with the additional assumptions of reasonably accurate inferences in small samples and insensitivity to sample size, but that combination of assumptions essentially amounts to our model.

A natural alternative modeling approach would be to build a theory of “sample-size neglect,” in which, loosely speaking, an agent forms beliefs about a sample of any size as if it were a “medium-sized” sample of, say, size 7. Such a model would imply under-inference for sample sizes larger than 7 and over-inference for sample sizes smaller than 7. This is a common conceptualization and one which we found compelling enough to consider as our first (and more parsimonious) approach. But there are two reasons why we settled on explaining the data with a combination of N BLLN and LSN instead. First, the approach here makes different predictions regarding inferences an agent draws from a sample of size 1. Because both our model of N BLLN and LSN predict that an agent’s beliefs are correct for a sample of size 1, the combined model in Appendix A also makes this prediction. A model of sample-size neglect, by contrast, must predict that over-inference is most extreme for a sample of size 1, a prediction that is not borne out by the evidence we review in Appendix B. Second, a model of sample-size neglect is neither psychologically nor mathematically rich enough to describe people’s beliefs about sequences of signals.⁴⁵

Our model of N BLLN is defined only when the signals are *i.i.d.* and binomial. There are some natural approaches to modeling N BLLN for non-*i.i.d.* signal sequences. Consider a binomial random process defined by a mapping from any initial rate, θ_0 , and any history of t observed signals, h_t , into a rate that the $(t + 1)^{\text{st}}$ signal will be an a -signal, $\theta(\theta_0, h_t)$. When Barney knows the initial rate is θ_0 , he forms his beliefs as if the initial rate were β , a random variable drawn from distribution $f_{\beta|\Theta}^{\psi}(\beta|\theta_0)$. For the first signal in a group, he believes that the probability of an a -signal is β , and for the $(t + 1)^{\text{th}}$ signal within that group, he believes that the probability of an a -signal is $\theta(\beta, h_t)$. This modeling approach can be applied not only when the signals truly are

⁴⁵The first concern could be addressed by a model of “sample-size misperception,” in which an agent correctly perceives a sample of size 1, treats a small sample as if it were larger than it is, but treats all large samples as a sample of size 7. Such a model is similar in spirit to a combination of N BLLN and LSN but does not address the second concern.

non-*i.i.d.*, but also when an agent falsely believes they are non-*i.i.d.* due to another psychological bias. Indeed, Appendix A applies this approach to combine NBLLN with belief in the gambler's fallacy, which posits that people think that signals are negatively autocorrelated even when they are *i.i.d.*.

There are also natural extensions of our modeling approach to non-binomial cases.⁴⁶ Suppose, for example, that the signals are normally distributed *i.i.d.* with known mean μ and variance σ^2 . We can imagine a cousin of Barney believes instead that signals are generated by a two-stage process, where a subjective mean ν is drawn from some distribution centered at μ , and then the signals are drawn from a normal distribution with mean ν and variance σ^2 . While Tommy believes that the mean of a large random sample of signals will converge to a point mass at μ , Barney's cousin believes it will converge to the density of ν . We could assume that the density of ν corresponds to the empirical large-sample beliefs, or for analytical tractability, we could assume that ν follows the conjugate prior distribution for the normal distribution, which is itself a normal distribution.

⁴⁶While experiments have predominantly focused on the binomial case, in Appendix B we discuss some evidence that underinference in large samples applies also to multinomial and normal signals.

8 Appendix A: A Combined Model of Non-Bayesian Updating

In this appendix, we briefly outline models of other biases besides NBLLN that also appear to matter for inference, focusing on the same single-clump context of Section 2. These models help organize our review of the evidence in Appendix B. Although these models are far more cursory and preliminary than our model of NBLLN, we hope that formalizing these alternatives can both crisply differentiate them from NBLLN—clarifying in particular that none of them are the “opposite” of or inconsistent with NBLLN—and also be suggestive of how to develop these other biases along lines we have done with NBLLN.

At first glance, NBLLN appears to be directly at odds with another bias in beliefs, the Law of Small Numbers (LSN): Tversky & Kahneman (1971) formulated the term “Law of Small Numbers” to refer to the idea that people exaggerate the likelihood that small samples will reflect the underlying population. While NBLLN generally leads to under-inference, LSN generates over-inference. However, these two biases are neither logically nor psychologically inconsistent. Indeed, we believe it is the combination of the two which has led judgment researchers to posit a bias of “sample-size neglect” in which people overestimate the resemblance of small samples and underestimate the resemblance of large samples as if they simply do not see the relevance of sample size. We re-interpret what appears to be sample-size neglect as a combination of these two biases, and we show how the basic under-inference implications of NBLLN goes through after LSN is accounted for.

According to LSN, people exaggerate how much small samples resemble the population. In Rabin’s (2002) model, an agent forms beliefs about N *i.i.d.* draws that have known rate θ as if signals were drawn without replacement from an “urn” of size M that contains exactly θM a -signals and $(1 - \theta)M$ b -signals.⁴⁷ To make sure that the agent continues to view the draws as random even after many signals have been realized, it can be assumed that $M > 2$, and the urn is “renewed” every odd number of draws. In other words, every odd-numbered draw, the signal is drawn from a refilled urn of size M , and every even-numbered draw, the signal is drawn from the urn of size $M - 1$ that is depleted by the previous signal’s draw. An agent who believes in LSN is called Freddy. The parameter M governs the strength of LSN, with Freddy becoming Tommy in the parameter limit $M \rightarrow \infty$. In Rabin’s model, for a given θ , the parameter M , in addition to being an integer larger than 2, must satisfy the constraint that θM is an integer. This constraint becomes problematic when combining the model of LSN with our model of NBLLN. In our model of NBLLN, the agent thinks the outcome of a random sample is determined by the subjective rate

⁴⁷Rabin & Vayanos (2010) improves on Rabin’s (2002) model that we discuss here, generalizing LSN beyond the binomial case and without assuming the signal-generating process is *i.i.d.* We explore this more contrived simple model of LSN here for ease of combining it with NBLLN.

β , which is drawn from a distribution with full support on $[0, 1]$. Hence for any M , βM will be non-integer-valued with probability 1. For this reason, we propose a variant of Rabin’s model that, while still requiring that M is an integer larger than 2, remains well-defined even when θM is not an integer.

Our variant of Rabin’s model of LSN is identical to Rabin’s model except that instead of believing that the “urn” contains θM a -signals and $(1 - \theta) M$ b -signals, Freddy believes that it contains \tilde{A} a -signals and $(M - \tilde{A})$ b -signals, where \tilde{A} is an integer-valued random variable that equals $j \in \{0, 1, \dots, M\}$ with probability $\binom{M}{j} \theta^j (1 - \theta)^{M-j}$. In words, Freddy thinks that signals are drawn without replacement from an urn of size M , but he believes the composition of the urn is random, with the “average” urn containing θM a -signals and $(1 - \theta) M$ b -signals. When the rate is known to be θ , Freddy believes that the number of a -signals in the urn is a binomial random variable with parameters (θ, M) .

NBLLN is the belief that the mean of a random sample converges to a non-trivial distribution, rather than a precise estimate of the mean of the population, in the limit as the sample size gets large. We refer to an agent who believes in both LSN and NBLLN as Barney-Freddy, with beliefs denoted by $f^{\psi M}$. Like Barney, he predicts that the sample is drawn according to a subjective rate that may not equal the true rate. We assume that the subjective rate $\beta \in [0, 1]$ is drawn from a density $f_{\beta|\Theta}^{\psi}(\beta|\theta)$ determined by θ and Barneyess parameter ψ that satisfies Assumptions A1-A4 from Section 2. In accordance with LSN, Barney-Freddy thinks that signals are drawn without replacement from an “urn” of size M that contains \tilde{A} a -signals and $(M - \tilde{A})$ b -signals, where \tilde{M} is a binomial random variable with parameters (β, M) . Hence Barney-Freddy thinks that the “average” urn contains βM a -signals and $(1 - \beta) M$ b -signals. The integer $M > 2$ parameterizes the degree of belief in LSN, and we assume that the urn is “renewed” and \tilde{A} is re-drawn every odd number of signals. As usual for the model of NBLLN, the same subjective rate β applies to the entire clump.

Barney-Freddy both believes in the gambler’s fallacy—that is, he expects recent a -signals to be followed by b ’s and vice-versa—and believes that the sample mean of a large population will converge toward a full-support limit distribution. He does not observe the subjective rate β , but conditional on any given β , Barney-Freddy expects that if the even-numbered draw t is more likely to be an a signal if the $(t - 1)^{st}$ draw was b than if it was a . Hence even without knowing β , Barney-Freddy excessively expects that a and b signals will alternate between an odd draw and an even draw. Consequently, for reasonable calibrations of ψ and M , Barney-Freddy’s subjective sampling distribution in a finite sample will be too peaked, putting too little weight in the tails. On the other hand, Lemma A1 states that Barney-Freddy’s beliefs about a large sample converge

to a full-support limit distribution. Like for Barney, Barney-Freddy's limit density will be equal to $f_{\mathbb{B}|\Theta}^\psi(\beta|\theta)$, regardless of the degree of belief in LSN. Intuitively, every odd-even pair of draws will have, on average, proportion β of a -signals. Hence by the Law of Large Numbers, the sample as a whole will tend toward having proportion β of a -signals almost surely.

Lemma A1. *Barney-Freddy does not believe in LLN: For any interval $[\alpha_1, \alpha_2] \subseteq [0, 1]$,*

$$\lim_{N \rightarrow \infty} \sum_{x=[\alpha_1 N]}^{[\alpha_2 N]} f_{S_N|\Theta}^{\psi M}(A_s = x|\theta) = F_{\mathbb{B}|\Theta}^\psi(\beta = \alpha_2|\theta) - F_{\mathbb{B}|\Theta}^\psi(\beta = \alpha_1|\theta) > 0.$$

Not only are LSN and NBLLN mutually consistent, but LSN actually *magnifies* NBLLN. For smaller M , Barney-Freddy believes that odd-even signal pairs more frequently alternate, and hence his subjective sampling distribution converges to the limit distribution more quickly.

While it is true that Freddiness generates over-inference while Barneyeness tends to generate under-inference, there is a clear pattern to when Barney-Freddy over-infers and when he under-infers. For reasonable calibrated values of ψ and M , Barney-Freddy will over-infer from small samples. For any values of ψ and M , it follows immediately from Lemma A1 that Barney-Freddy will under-infer when the sample size N is sufficiently large.

A well-known bias is base-rate neglect (Kahneman & Tversky, 1973), an underweighting of prior probabilities in drawing inferences. Instead of assuming that an agent updates according to Bayes's Rule applied to his subjective sampling distributions, we can capture base-rate neglect by assuming that the agent draws inferences according to:

$$f_{S_N|\Theta}^{\psi Mb}(\theta_A|s) = \frac{f_{S_N|\Theta}^{\psi M}(s|\theta_A) f_\Theta(\theta_A)^b}{f_{S_N|\Theta}^{\psi\phi\gamma}(s|\theta_A) f_\Theta(\theta_A)^b + f_{S_N|\Theta}^{\psi\phi\gamma}(s|\theta_B) f_\Theta(\theta_B)^b}$$

and

$$f_{S_N|\Theta}^{\psi Mb}(\theta_B|s) = \frac{f_{S_N|\Theta}^{\psi M}(s|\theta_B) f_\Theta(\theta_B)^b}{f_{S_N|\Theta}^{\psi M}(s|\theta_A) f_\Theta(\theta_A)^b + f_{S_N|\Theta}^{\psi M}(s|\theta_B) f_\Theta(\theta_B)^b},$$

where $f_{S_N|\Theta}^{\psi M}(s|\theta_A)$ and $f_{S_N|\Theta}^{\psi M}(s|\theta_B)$ are the subjective sampling distributions of an agent with Barneyeness parameter $0 < \psi < \infty$ and Freddiness parameter $M > 2$, and $0 \leq b \leq 1$ parameterizes the degree of base-rate neglect. If $b = 1$, these formulae specialize to Bayes' Rule, where there is no base-rate neglect. If $b = 0$, the agent ignores base rates altogether, treating any prior probabilities as if they were 50-50. This formulation of base-rate neglect has been previously adopted in empirical work by (e.g., Grether, 1980) and concurrently in theoretical work by Bodoh-

Creed (2010). Applying it theoretically in dynamic settings raises many of the same conceptual issues as NBLLN; the way in which the agent processes groups of signals will matter a great deal in how his beliefs evolve. For that reason, we believe the framework we have begun to develop in this paper for analyzing dynamic NBLLN may be of use for analyzing dynamic base-rate neglect as well.

A simple explanation for some of the evidence on people’s beliefs is “extreme-belief aversion,” an aversion to holding beliefs that are close to certainty. Consider a discrete probability density function, $f_X(\cdot)$, that puts positive probability on a set of possible outcomes x_1, x_2, \dots, x_J . We capture the idea of extreme-belief aversion by defining a mapping from the true probability density, $f_X(\cdot)$, to a subjective probability density that is less extreme,

$$f_X^\xi(x_i) = \frac{.5 + \xi \cdot (f_X(x_i) - .5)}{\sum_{j=1}^J (.5 + \xi \cdot (f_X(x_j) - .5))}.$$

The parameter $0 < \xi < 1$ describes the degree of extreme-belief aversion, with smaller values corresponding to greater bias. If $\xi = 1$, the subjective probabilities coincide with the true probabilities, while if $\xi = 0$, all outcomes x_1, x_2, \dots, x_J are treated as equally likely.

One interpretation of the transformed probability, $f_X^\xi(x)$, is that it represents the agent’s truly-held beliefs. Another interpretation is that the agent actually holds beliefs $f_X(x)$ but *reports* beliefs that are transformed by the ξ function. While the latter is certainly plausible when Tommy’s beliefs are very extreme—we can easily imagine a person saying she is 99% sure when her true belief is .9999—it does not address the evidence that beliefs inferred from betting behavior also exhibit under-inference.

While extreme-belief aversion does seem to help describe the evidence on people’s beliefs, and it may be a confound for other interpretations of biased beliefs, we do not review evidence for extreme-belief aversion, and we know of no evidence that our crude formulation is a close match for people’s thinking. Extreme-belief aversion may also lead to internal-inconsistency modeling challenges that we do not address. For example, it seems reasonable to assume that the transformation above could be applied to an agent’s sampling distribution or to the agent’s inferences, depending on which beliefs are being elicited. In that case, however, subjective sampling distributions and inferences will not in general be linked by Bayes’ Rule.

Combining the three biases above with NBLLN gives a complicated model that captures many features that could be applied to predict beliefs and behavior in economic settings. One insight that comes immediately out of the combination is that base-rate neglect—i.e., underweighting priors—is not the opposite of, or contradictory to, the way NBLLN leads people to underweight likelihood

information. Indeed, as we have noted in some discussions above about vividness and other biases, NBLLN is, especially in understanding “multi-clump” information processing, likely a contributor to the relevance of other biases. In many information-rich environments where full Bayesians would correctly become very confident independent of their priors, NBLLN is necessary for the question of whether people neglect base rates to be relevant. At the same time, we show above why NBLLN means that unless people completely neglect base rates, people’s initial beliefs matter even in the long run. In combination, in fact, NBLLN suggests that it is possible that the real and often important fact that people under-use base rates may be consistent with the possibility that base rates matter more for social and economic phenomena than fully rational models have supposed that they do.

One last bias sits less comfortably with the others, and is harder to integrate. Many experiments make clear that peoples’ subjective sampling distributions have flatter tails than our model of NBLLN by itself can explain. We attribute the flatness to “sampling-distribution-tails diminishing sensitivity (SDTDS),” a bias in which people perceive very unlikely outcomes as similar to each other and hence similar in probability. Consider 10 flips of a coin that is biased .8 in favor of heads. We know of no direct evidence but conjecture that most people would judge the likelihood of observing 1 head as very close to the likelihood of observing 0 heads, even though observing 1 head is actually 40 times more likely.

SDTDS can be formalized by assuming that an agent forms beliefs as if the likelihood of sample realizations far from the average are more similar to each other (and to the average) than they truly are. For a sample of size N from a θ -biased coin, let σ denote the standard deviation of the sample proportion $\frac{A_N}{N}$ (which equals $\sqrt{\frac{\theta(1-\theta)}{N}}$ for Tommy but not for Barney-Freddy). Let the perceived distance between the realized $\frac{A_N}{N}$ and θ be $\sigma\gamma\left(\frac{\frac{A_N}{N}-\theta}{\sigma}\right)$, where the twice-differentiable “sample-perception function” $\gamma : (-1, 1) \rightarrow (-1, 1)$ has the properties (1) $\gamma(0) = 0$, (2) $0 < \gamma' < 1$, (3) $\gamma''(x) < 0$ for all $x > 0$, and $\gamma''(x) > 0$ for all $x < 0$, and (4) $\gamma(x) = -\gamma(-x)$ for all x .

Property (1) says that the agent perceives a sample proportion of θ accurately. Property (2) ensures that the agent perceives any other sample proportion as more similar to θ than it actually is. The key concavity and convexity assumption (3) means that neighboring samples are perceived as more similar to each other the further they are from θ . For this reason, the agent makes little distinction between outcomes that fall far in the tails of his subjective sampling distribution. Property (4) specifies that γ is symmetric around 0 so that the misperception is symmetric around θ .

Roughly speaking, a person who exhibits SDTDS with sample-perception function γ judges the probability of a sample $s \in S_N$ with proportion $\frac{A_N}{N}$ of a -signals as if it were the sample

$\Lambda(s)$, which has proportion $\theta + \sigma\gamma \left(\frac{A_N - \theta}{\sigma} \right)$ a -signals. Formally, $f_{S_N|\Theta}^{\psi\phi\gamma}(s|\theta) = \frac{f_{S_N|\Theta}^{\psi\phi}(\Lambda(s)|\theta)}{\sum_{s' \in S_N} f_{S_N|\Theta}^{\psi\phi}(\Lambda(s')|\theta)}$, where $f_{S_N|\Theta}^{\psi\phi}$ is the subjective sampling distribution for $\psi\phi$ -Barney-Freddy, and the denominator is a normalization that ensures that the subjective sampling distribution adds up to 1.⁴⁸ While NBLLN by itself leads to subjective sampling distributions that have flat tails, SDTDS is an additional force for flat tails. When the rate θ is not .5, the mean of the agent’s subjective sampling distribution no longer equals the mean of the objective sampling distribution because the “long tail” of the distribution is overweighted. Both of these features—flatter tails than can be accommodated with a reasonably-calibrated model of NBLLN and a mean shifted toward the long tail—are present in the subjective sampling distributions measured by Kahneman & Tversky (1972).

Unfortunately, this formulation of SDTDS cannot be so easily integrated with NBLLN—or with a model combining NBLLN with LSN and base-rate neglect—because it does not arm somebody with a theory of the sequence of signals, only the frequency of signals within a sample. While surely a form of it could be specified that embeds a concrete theory of permutations within the sample (allowing for instance a person to believe all different sequences are equally likely), we do not know that the existing evidence provides a guide, nor do we believe the psychology underlying it translates easily into situations where an agent is cognizant of the ordering of signals. Moreover, an improved model or a better formulation than we have found to interpret existing evidence may be more compatible with the models of other biases than we have supposed.

⁴⁸In general, $\Lambda(s)$ could be a sample outside the support of the objective sampling distribution. For example, it may be a sample with 3.2 a ’s. The expression above is nonetheless well-defined as long as the density $f_{S_N|\Theta}^{\psi\phi}(\cdot|\theta)$ can be evaluated at that sample, as it can for a binomial objective sampling distribution.

9 Appendix B: Experimental Evidence

In this appendix, we report on all papers we could identify with experimental results related to the model’s assumptions about subjective sampling distributions and predictions about inference for binomial signals. Table B1 lists the papers, which assumptions or predictions they test, their experimental subject population, and their incentive structure. Most of the studies we review did not incentivize subjects’ responses; we will also discuss how the evidence from the few incentivized experiments relates to the unincentivized studies.

Table B1. Experimental evidence on NBLLN for binomial signals.

Author(s)	Year	Prediction or Inference?	Subjects	Incentives?
Beach, Wise, & Barclay	1970	inference	169 male undergrads	no
Camerer	1987	inference	74 undergrads	financial market
Chinnis & Peterson	1968	inference	40 male undergrads	no
Dave & Wolfe	2003	inference	40 undergrads	BDM for probability
DeSwart	1972	inference	21 male undergrads	no
DeSwart	1972	inference	18 male undergrads	no
Donnell & DuCharme	1975	inference	24 male undergrads	no
Gettys & Manley: Study 1	1968	inference	20 undergrads	no
Gettys & Manley: Study 2	1968	inference	28 undergrads	no
Green, Halbert, and Robinson	1965	inference	32 grad students	paid for guess about state
Grether	1980	inference	341 undergrads	paid for guess about state
Grether: Studies 1 and 2	1992	inference	97 undergrads	paid for guess about state
Grether: Study 3	1992	inference	55 summer students	BDM for probability
Griffin & Tversky: Study 1	1992	inference	35 undergrads	paid for accurate posterior
Griffin & Tversky: Study 2	1992	inference	40 undergrads	paid for accurate posterior
Griffin & Tversky: Study 3	1992	inference	50 undergrads	no
Kahneman & Tversky: prediction	1972	prediction	unclear	no
Kahneman & Tversky: inference	1972	inference	560 high school students	no
Kraemer & Weber	2004	inference	51 students (most grad)	paid for accurate posterior
Marks & Clarkson	1972	inference	68 undergrads	no
Nelson, Bloomfield, Hales, & Libby: Study 1	2001	inference	27 MBA students	financial market
Peterson & Miller	1965	inference	42 undergrads	no
Peterson & Swensson: Study 1	1968	inference	15 male undergrads	no
Peterson & Swensson: Study 2	1968	inference	18 male undergrads	no
Peterson, DuCharme, & Edwards: Study 1	1968	prediction	41 male undergrads	no
Peterson, DuCharme, & Edwards: Study 2	1968	both	24 male undergrads	no
Peterson, Schneider, & Miller	1965	inference	44 undergrads	no

Author(s)	Year	Prediction or Inference?	Subjects	Incentives?
Phillips & Edwards: Study 1	1966	inference	5 male undergrads	no
Phillips & Edwards: Study 2	1966	inference	48 male undergrads	paid for accurate posterior
Phillips & Edwards: Study 3	1966	inference	48 male undergrads	no
Pitz	1967	inference	28 undergrads	no
Sanders	1968	inference	32 undergrads	bets on the state
Sasaki & Kawagoe	2007	inference	1033 employees	no
Strub	1969	inference	12 male undergrads	paid for guess about state
Teigen: Study 1	1974	prediction	22 undergrads	no
Teigen: Study 2	1974	prediction	73 undergrads	no
Wheeler & Beach	1968	both	17 male undergrads	paid for beliefs (prediction), and bets (inference)

9.1 Evidence on Subjective Sampling Distributions

We begin by assessing the model’s assumptions and predictions about subjective sampling distributions. We have found only 6 experiments from 4 papers in which researchers explicitly elicited experimental participants’ beliefs about the likelihood of each possible sample.⁴⁹ None of these elicitations were incentivized. For 5 of these studies—Kahneman & Tversky (1972), Peterson, DuCharme & Edwards’s (1968) Study 1, Wheeler & Beach (1968), and both of Teigen’s (1974) studies—the data are displayed in the paper, and we have reproduced the graphs in Figures 1, B4, B5, and B1, respectively (we display both of Teigen’s studies together), shown below in the order we discuss them.

Among the papers, Kahneman & Tversky (1972) elicited sample-proportion beliefs for the largest sample sizes. As discussed in the Introduction, they find that subjective sampling distributions are “constant in proportions” for $N = 10$, 100, and 1000. There is no noticeable tightening of the distribution even for $N = 1000$; while in fact there is less than a .01 chance of the proportion of heads falling outside the range 45% to 55%, subjects’ distributions assign probability .79 to a proportion outside that range.⁵⁰

A straightforward implication of the model is that subjective sampling distributions will be flatter than the objective sampling distributions. In all 7 experiments, with the exception of the $N = 3$ conditions of one experiment, the researchers indeed concluded that subjective sampling

⁴⁹Cohen and Hansel (1955) also elicited subjective sampling distributions, but we cannot compare their data with our model because they did not tell their subjects the rate that was generating the signals.

⁵⁰At the time of this draft, two of the authors (Benjamin and Rabin, in joint work with Don Moore) have also collected data on people’s subjective sampling distributions for $N = 10$ and 1000. We designed the experiment to measure subjective sampling distributions in several different ways to deal with potential confounds such as extreme-belief aversion. When we elicit distributions in the same manner as Kahneman & Tversky, we replicate their results almost exactly. While our preliminary findings support NBLLN for $N = 1000$, we also find that the evidence for NBLLN for $N = 10$ —and hence presumably also the evidence about smaller sample sizes reviewed below—is confounded by other explanations.

distributions are excessively close to uniform.⁵¹

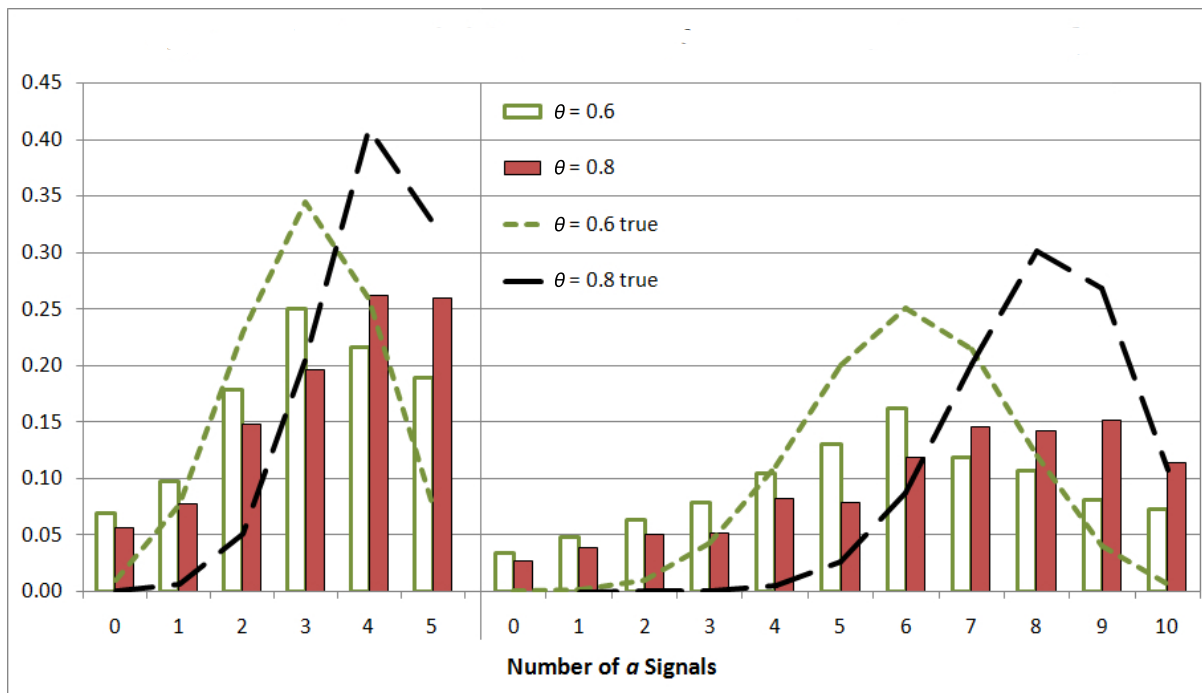


Figure B1: Median probability estimates, $N=5$ and $N=10$ (Teigen, 1974)

A feature of our model of NBLLN is that people have correct beliefs about samples of size 1. We know of no evidence on this point, but our own introspection suggests it is virtually a tautology that if a person knows the rate of a -signals is θ , the person thinks the probability is θ that a single draw will be a .⁵² However, our model of NBLLN implies that for any sample size larger than 1, no matter how small, the subjective sampling distribution will be too close to uniform. Peterson, DuCharme & Edwards's (1968) Study 1 $N = 3$ conditions are the one case where researchers found subjective sampling distributions that are *not* too flat. For all three rates they studied ($\theta = .6, .7, .8$), the $N = 3$ subjective sampling distribution nearly coincides with the objective sampling distribution. We interpret this evidence as consistent with the combined effects of NBLLN and LSN, rather than either of the biases considered alone. The combined model we discuss in Appendix A predicts that subjective sampling distributions will be correct for $N = 1$ and too flat for N large. For plausible parameter values, there will be a non-monotonicity for intermediate N : due to LSN, the subjective

⁵¹Unlike the other 6 studies, Teigen (1974) asked subjects about the probability of each possible outcome separately, without requiring that the probabilities sum to 1. He found that the probabilities summed to greater than 1, and subjects often assigned probabilities that were too high to every outcome. When the subjective sampling distributions were normalized to sum to 1, they were too flat.

⁵²Of course there is a good reason no one has done this experiment: the correct answer about the probability of an a -signal is obvious if the experimenter has just said that it is θ . Nonetheless, we discuss below evidence from experiments on inference from a single signal. The correct answer to an inference problem is not obvious, and indeed subjects often do not get the correct answer, tending somewhat to under-infer on average.

sampling distribution may be too peaked when N is larger than 1 but small. The relative strength of NLLN grows with N , so the subjective sampling distribution may be nearly correct for some intermediate N .

Assumption A4 of our model is that the subjective sampling distribution has the same mean as the objective sampling distribution, regardless of sample size. In contrast, the evidence indicates that when $\theta \neq .5$, the mean of the subjective sampling distribution is generally between the objective mean, θN , and $.5$. Kahneman & Tversky (1972) explicitly comment that “the mean is displaced towards the long tail” (p.440), and this pattern is visually evident in all of the figures except the $N = 3$ cases discussed above. From a modeling perspective, we believe it is appropriate for our model to have the counterfactual feature of an accurate mean for the subjective sampling distribution because it allows us to draw out the implications of believing that the limit distribution has full support, without mixing in the implications of an inaccurate mean. Moreover, we speculate that the “displaced mean” is the result of a psychologically distinct bias, “sampling-distribution-tails diminishing sensitivity (SDTSD),” sketched in Appendix A, and our calibrated model of NLLN generates subjective sampling distributions that come closer to matching the data in Figure 1 when we additionally incorporate SDTSD into the model (calculations not shown).

There is mixed evidence about whether training and feedback affects subjective sampling distributions. Wheeler & Beach (1968) elicited subjective sampling distributions for a sample of size $N = 8$ for rates $\theta = .6$ and $.8$; these are shown in Figure B5. Next, their subjects were faced with 100 asymmetric binomial inference problems ($\theta_A = .8$ and $\theta_B = .4$). After each problem, the subject was told the true rate for that problem. Subjective sampling distributions were elicited again, the subjects responded to 100 more symmetric binomial inference problems with feedback, and the subjective sampling distributions were elicited a final time.⁵³ Comparing the initial subjective sampling distributions with the final ones, the final subjective sampling distributions are less flat. For $\theta = .8$, the final subjective sampling distribution is quite close to the objective sampling distribution. For $\theta = .6$, the final subjective sampling distribution is actually more peaked than the objective sampling distribution. On the other hand, Peterson, DuCharme, and Edwards (1968, Study 2) conducted an experiment with four stages: (1) subjects were faced with symmetric binomial inference problems (with no feedback); (2) subjective sampling distributions were elicited for each combination of $N = 3, 5, 8$ and $\theta = .6, .7, .8$; (3) subjects were shown the objective sampling distributions; and (4) subjects were faced with another series of symmetric binomial inference problems. Subjects’ responses in the inference problems were similar in stage 4 as in stage 1, suggesting that showing subjects the objective sampling distributions had little effect on beliefs.

⁵³These subsequent elicitations are not shown in Figure B5.

9.2 Evidence on Inference

We have found 33 studies from 26 papers measuring inferences from samples about which of two equally-likely rates, θ_A and θ_B , generated the samples. Most of these binomial inference problems are symmetric in the sense that $\theta_A = 1 - \theta_B$. Unless otherwise noted, all the studies we mention are symmetric. We focus first on the studies where the prior probabilities, $f_{\Theta}(\theta_A)$ and $f_{\Theta}(\theta_B)$, are equal. Equal priors neutralizes the role of base-rate neglect. We study below how inferences are affected by unequal priors.

To compare the degree of under- or over-inference across studies, note that Bayes' Rule can be written as $\frac{f_{\Theta|S_N}(\theta_A|s)}{f_{\Theta|S_N}(\theta_B|s)} = \frac{f_{S_N|\Theta}(s|\theta_A)f_{\Theta}(\theta_A)}{f_{S_N|\Theta}(s|\theta_B)f_{\Theta}(\theta_B)}$. Since the signals are binomial, the rates are symmetric, and the priors are equal, Bayes' Rule can be expressed as $\frac{f_{\Theta|S_N}(\theta_A|s)}{f_{\Theta|S_N}(\theta_B|s)} = \left(\frac{\theta_A}{1-\theta_A}\right)^{\frac{2A_s-N}{N} \times N}$. Taking the natural log twice and rearranging,

$$\ln \ln \frac{f_{\Theta|S_N}(\theta_A|s)}{f_{\Theta|S_N}(\theta_B|s)} - \ln \left(\frac{2A_s - N}{N}\right) - \ln \ln \left(\frac{\theta_A}{1 - \theta_A}\right) = \ln N. \quad (5)$$

It is possible in 9 of the papers to identify the value of θ_A for the inference problem, the actual sample observed by subjects, and subjects' mean or median reported posterior. Using the experimental data, Figure B2 plots the left-hand side of equation (5) against $\ln N$.⁵⁴ If the subjects' beliefs were Bayesian, the points should cluster along the identity line (the dashed line in Figure B2). The best-fitting regression line (the solid line in the figure) has a slope smaller than 1, indicating that subjects generally infer less in favor of rate θ_A than a Bayesian would.

To estimate the degree of under-inference and to probe its robustness, we rewrite (5) as a regression equation:

$$\ln \ln \frac{f_{\Theta|S_N}(\theta_A|s)}{f_{\Theta|S_N}(\theta_B|s)} = \gamma_0 + \gamma_1 \ln N + \gamma_2 \ln \left(\frac{2A_s - N}{N}\right) + \gamma_3 \ln \ln \left(\frac{\theta_A}{1 - \theta_A}\right) + \varepsilon. \quad (6)$$

⁵⁴The left-hand side is well-defined only for inference problems such that $\frac{\theta_A}{1-\theta_A} > 1$ (that is, $\theta_A > .5$) and $\frac{A_s}{N} > \frac{1}{2}$ (that is, over half the realized signals are a 's). Hence, as written, equation (5) only applies to such cases. Although this holds for only 63 of the 99 regression observations in Figure B2 and Table B2, we can include additional regression observations by relabeling the rates and sample proportions that we plug into formula (5). In inference problems such that $\frac{\theta_A}{1-\theta_A} > 1$ and $\frac{A_s}{N} < \frac{1}{2}$, we express Bayes' Rule as $\frac{f_{\Theta|S_N}(\theta_B|s)}{f_{\Theta|S_N}(\theta_A|s)} = \left(\frac{\theta_B}{1-\theta_B}\right)^{\frac{2A_s-N}{N} \times N} = \left(\frac{\theta_A}{1-\theta_A}\right)^{\frac{N-2A_s}{N} \times N}$, so equation (5) becomes $\ln \ln \frac{f_{\Theta|S_N}(\theta_A|s)}{f_{\Theta|S_N}(\theta_B|s)} - \ln \left(\frac{N-2A_s}{N}\right) - \ln \ln \left(\frac{\theta_A}{1-\theta_A}\right) = \ln N$. This allows us to use an additional 32 regression observations. Finally, we can use a further 4 regression observations for which $\theta_A < .5$ and $\frac{A_s}{N} < \frac{1}{2}$ by expressing Bayes' Rule as $\frac{f_{\Theta|S_N}(\theta_A|s)}{f_{\Theta|S_N}(\theta_B|s)} = \left(\frac{\theta_A}{1-\theta_A}\right)^{\frac{2A_s-N}{N} \times N} = \left(\frac{1-\theta_A}{\theta_A}\right)^{\frac{N-2A_s}{N} \times N}$, and we can take the log-log of this equation. In the case of 3 inference problems, $\frac{A_s}{N} = \frac{1}{2}$, so the Bayesian posterior ratio is equal to 1, and it is impossible to define what constitutes "over-inference" or "under-inference." Those 3 datapoints are dropped from the Figure B2 and Table B2.

The null hypothesis of Bayesian updating is $\gamma_0 = 0$, $\gamma_1 = \gamma_2 = \gamma_3 = 1$. In Table B2, we estimate versions of equation (6) with several different restrictions on the coefficients and data.

Table B2. Inference with symmetric rates and equal priors.

	(1)	(2)	(3)	(4)	(5)
	Restriction:	Restriction:	Restriction:	Coeffs	Coeffs unrestricted
	$\gamma_1 = \gamma_2 = \gamma_3 = 1$	$\gamma_2 = \gamma_3 = 1$	$\gamma_3 = 1$	unrestricted	Incentivized only
$\ln N$		0.587 (0.083)	0.470 (0.109)	0.485 (0.085)	0.675 (0.086)
$\ln \left(\frac{2A_s - N}{N} \right)$			0.776 (0.116)	0.914 (0.109)	0.957 (0.121)
$\ln \ln \left(\frac{\theta_A}{1 - \theta_A} \right)$				0.333 (0.101)	0.421 (0.104)
Constant	-0.753 (0.076)	-0.031 (0.143)	-0.002 (0.142)	-0.145 (0.108)	-0.168 (0.130)
R^2	0.000	0.426	0.360	0.603	0.763
#obs	99	99	99	99	47
#papers	9	9	9	9	3

Notes: Results are from OLS regressions, with standard errors in parentheses. The dependent variable is as described in the text. Coefficients for blank entries are restricted to equal 1. The fifth column restricts the data to incentivized experiments.

Column 1 estimates just γ_0 , under the restriction that $\gamma_1 = \gamma_2 = \gamma_3 = 1$. The estimate, $\hat{\gamma}_0 = -0.753$, is significantly smaller than zero, indicating that the average pattern is under-inference.

Column 2 estimates both γ_0 and γ_1 , while restricting $\gamma_2 = \gamma_3 = 1$. The predicted regression line is plotted as the solid line in Figure B2. The regression confirms that the degree of under-inference is related to sample size; $\hat{\gamma}_1 = .587$ is significantly smaller than 1 (but greater than 0). Moreover, once the degree of under-inference is allowed to depend on sample size, there is no residual under-inference to be picked up by the constant term: $\hat{\gamma}_0 = -0.031$, which is not statistically distinguishable from zero. Breaking down the data by study (not shown in the table), every study that manipulates sample size, while holding constant other features of the experiment—Peterson, Schneider, and Miller (1965); Pitz (1967); Peterson, DuCharme, & Edwards’s (1968) Study 2;

Kahneman & Tversky (1972); Griffin & Tversky’s (1992) Study 1; Nelson, Bloomfield, Hales, & Libby’s (2001) Study 1; Kraemer & Weber (2004)—concludes that there is greater under-inference in larger samples.⁵⁵

Qualitatively, extreme-belief aversion can explain both the excessively-dispersed subjective sampling distributions discussed in section B.1 and the under-inference from column 1. An agent with extreme-belief aversion will have a more dispersed subjective sampling distribution than Tommy in large samples by virtue of compressing beliefs away from 0 and 1. An agent with extreme-belief aversion will also under-infer in situations where Tommy’s inference would be extreme, as would almost always occur when the sample is large. Extreme-belief aversion taken alone, however, implies that in any two inference problems where Tommy’s posterior is the same, people would hold (or at least report) the same belief. There is evidence contradicting this implication, indicating that extreme-belief aversion is not the *only* deviation from Bayesian belief formation that is going on.

For example, consider the experiment in Griffin & Tversky (1992), where $\theta_A = .6$. Tommy’s inference depends only on the difference between the number of *a*-signals and the number of *b*-signals. However, when the sample is 4 *a*’s and 1 *b*, subjects’ median belief in favor of θ_A is .80, while when the sample is 10 *a*’s and 7 *b*’s, subjects’ median belief in favor of θ_A is .60. Tommy’s belief would be .77 in both cases, so people are under-inferring from the sample of size 17 and actually slightly over-inferring from the sample of size 5. Consistent with Proposition 5—but inconsistent with extreme-belief aversion being the only bias in beliefs—people infer less from the same difference in *a* and *b* signals when the sample is larger. Kahneman & Tversky (1972) and Kraemer & Weber (2004) also report evidence that beliefs are sensitive to sample size, holding constant the difference in the number of *a* and *b* signals.

Column 3 of Table B2 relaxes the regression model still further, estimating γ_0 , γ_1 , and γ_2 , with the only remaining restriction being $\gamma_3 = 1$. The coefficient on the proportion of *a* signals, $\hat{\gamma}_2 = .776$, is just statistically distinguishable from 1, and the estimates $\hat{\gamma}_0$ and $\hat{\gamma}_1$ are not substantially different in column 3 compared with column 2.

Column 4 estimates all four coefficients. Qualitatively, the main effect of relaxing the $\gamma_3 = 1$ restriction on the conclusions from column 3 is that the coefficient on the proportion of *a*-signals, $\hat{\gamma}_2 = .914$, is no longer statistically distinguishable from 1. Hence we cannot reject the Bayesian null hypotheses that $\gamma_0 = 0$ and $\gamma_2 = 1$, while $\hat{\gamma}_1$ is significantly smaller than 1. However, column 4 makes clear that the extremeness of the rates—the extent to which $\theta_A = 1 - \theta_B$ departs from

⁵⁵For Green, Halbert, & Robinson (1965), we can also reach this conclusion by estimating the regression equation (6) on the data reported just in that paper. There are a number of other studies that manipulate sample size but do not analyze or display the data in a way that makes it clear how sample size affects the degree of bias in inference: Sanders (1968); Peterson & Swensson’s (1968) Study 2; Beach, Wise, & Barclay (1970); Marks & Clarkson (1972); and DeSwart (1972a, 1972b).

.5—also matters for the degree of biased inference. The coefficient $\hat{\gamma}_3 = .333$ is much smaller than 1. This means that subjects under-infer by more the further is θ_A from .5. Breaking down the data by study (not shown in the table), every study that manipulates θ_A , while holding constant other features of the experiment either concludes that there is greater under-inference for θ_A further from .5 (Green, Halbert, & Robinson, 1965; Phillips & Edwards’s, 1966 Study 1 and 3; Peterson & Miller, 1965; Peterson & Swensson’s, 1968 Studies 1 and 2; Peterson, DuCharme, & Edwards’s, 1968 Study 2; Sanders, 1968; Donnell & DuCharme, 1975; Kahneman & Tversky, 1972) or finds it without explicitly stating it (Chinnis & Peterson, 1968; Beach, Wise, & Barclay, 1970; Shu & Wu, 2003).⁵⁶ Griffin & Tversky (1992) use the term “discriminability” to describe the phenomenon of under-inference becoming more severe when θ_A and θ_B are further apart. In Griffin & Tversky’s (1992; Study 3) particularly clear evidence from asymmetric inference problems, subjects were asked to infer the likelihood of rate θ_A where the rates have equal priors, the sample has size $N = 12$, and number of a -signals is $A_s = 7, 8, 9$, or 10 . When the rates are close together, $(\theta_A, \theta_B) = (.6, .5)$, the subjects exhibit slight over-inference: a Bayesian’s posteriors for these four inference problems would be .54, .64, .72, and .80, respectively, while subjects’ median posteriors were .55, .66, .75, and .85. When the rates are further apart, $(\theta_A, \theta_B) = (.6, .25)$, subjects exhibited massive under-inference: whereas a Bayesian’s posteriors in these problems would be .95, .98, .998, and .999, respectively, subjects’ posteriors were .60, .70, .80, and .90. Griffin & Tversky’s evidence cannot be fully explained by extreme-belief aversion (that simply maps an objective posterior into a less extreme subjective one) because, for example, subjects’ posterior of .80 is identical whether $\theta_B = .5$ and 10 a -signals were observed or $\theta_B = .25$ and 9 a -signals were observed, but the objective posteriors are quite different in those two cases. Our interpretation of Griffin & Tversky’s evidence is that realized samples far in the tails of the subjective sampling generate extreme under-inference because the tails are excessively flat. While NLLN generates flat tails qualitatively, our one-parameter model of NLLN cannot fit this evidence quantitatively at plausible calibrated values of ψ . We believe SDTSD captures the right psychology for explaining flat tails and likely explains both Griffin & Tversky’s discriminability evidence and the evidence from a number of other studies that find $\gamma_3 < 1$.

Column 5 estimates the same regression as column 4, but with the data restricted to incentivized experiments. There are only three such studies—Green, Halbert, & Robinson (1965), Nelson, Bloomfield, Hales, & Libby (2001), and Kraemer & Weber (2004)—but the results in column 5 are largely similar to column 4, except that both $\hat{\gamma}_1 = .675$ and $\hat{\gamma}_3 = .421$ are larger, suggesting greater sensitivity to sample size and to rates when accurate inferences are rewarded. Nonetheless, both

⁵⁶DeSwart (1972a, 1972b) manipulates how far θ_A is from .5 but does not analyze or display the data in a way that makes it clear how it affects the degree of under-inference.

coefficients remain far less than 1, indicating substantial biases relative to Bayesian inference.

Training in inference appears to reduce but not eliminate under-inference. When subjects were told after each inference which state actually occurred, they became less biased over time but still under-inferred by the end of the experiment (Phillips & Edwards's, 1966, Study 2; Camerer, 1987). Strub (1969) found under-inference among subjects who had received 114 hours of lecture sessions, demonstrations, problem-solving sessions, and other training in dealing with probabilities, including prior participation in inference experiments. When subjects were told after each inference what the normatively correct inference is, they very quickly learned to report more extreme beliefs, but they do not seem to have learned to draw better inferences. While reporting more extreme beliefs led to more accurate beliefs in problems where their pre-training beliefs were not extreme enough, it led to less accurate beliefs in problems where their pre-training beliefs were accurate (Donnell & DuCharme, 1975).

A few studies have found over-inference. In all cases, N is relatively small, and θ_A is relatively close to .5.⁵⁷ Griffin & Tversky's Study 1 (1992; $\theta_A = .6$), which compared inference from samples of size 3, 5, 9, 17, and 33, found over-inference for $N = 3$ and 5 and under-inference for the others. Nelson, Bloomfield, Hales, & Libby's Study 1 (2001; $\theta_A = .6$) conducted an experimental asset market where payoffs depended on correct inferences from samples of size 3 and 17. Subjects under-inferred for $N = 17$ and over-inferred for $N = 3$. There is some evidence, though it is weak, that over-inference in favor of a particular state occurs when the realized sample exactly matches the expected sample in that state, a phenomenon that has been called "exact representativeness." In an experimental asset market, Camerer (1987; $N = 3$, $\theta_A = .67$) found that the price of a state-contingent asset that pays off if state A is true was too high—indicating over-inference in favor of state A —when the observed sample contained exactly 2 a -signals and 1 b -signal, and symmetrically over-inference in favor of state B when the sample proportions were reversed. In an experimental asset market with asymmetric rates of $\theta_A = .67$ and $\theta_B = .5$, Grether (1980; $N = 6$) similarly found evidence indicating over-inference in favor of state A when the realized sample was 4 a 's and 2 b 's and over-inference in favor of state B when the realized sample was 3 a 's and 3 b 's. In a similar experiment, Grether (1992) found less support for "exact representativeness." Neither NLLN alone, nor NLLN combined with SDTSD, can explain over-inference.

Such over-inference can, however, be explained by the Law of Small Numbers (LSN). As shown in Appendix A, when NLLN and LSN are combined in a single model, small N is a necessary condition for over-inference. Moreover, we have argued that SDTSD will tend to generate under-

⁵⁷Peterson & Swenson's (1968) Study 1 finds over-inference for $N = 1$ and $\theta_A = .6, .67, .75, .9$ in the first half of their data. In the same inference problems in the second half of their data from Study 1, and in both halves from Study 2, however, they find under-inference.

inference when θ_A and θ_B are far apart.

A distinctive feature of our theory—a feature that differentiates it from alternative theories of under-inference from large samples discussed in Section 7 and Appendix A—is the prediction that inferences from a sample of size 1 will be correct. Sample-size neglect predicts over-inference for samples of size 1, while extreme-belief aversion predicts under-inference for samples of any size, including 1. There are 11 experiments that measure inference when $N = 1$. Peterson, Schneider, & Miller (1965; $\theta_A = .6$), Dave & Wolfe (2003; $\theta_A = .7$), Peterson & Swenson’s (1968; $\theta_A = .6, .67, .75, .9$) Study 2, and Gettys & Manley’s (1968) Studies 1 and 2 (which used a variety of asymmetric inference problems) found substantial under-inference. Chinnis & Peterson (1968; $\theta_A = .67, .8$), Kraemer & Weber (2004; $\theta_A = .6$), and Sasaki & Kawagoe (2007; $\theta_A = .67$) found slight under-inference, very close to Bayesian, and Peterson & Swenson’s (1968; $\theta_A = .6, .67, .75, .9$) Study 1 found over-inference in the first half of their data and under-inference in the second half. In a mix of symmetric and asymmetric problems, Peterson & Miller (1965) found under-inference for $(\theta_A, \theta_B) = (.83, .17)$, $(\theta_A, \theta_B) = (.71, .2)$, and $(\theta_A, \theta_B) = (.67, .33)$, and over-inference for $(\theta_A, \theta_B) = (.6, .43)$. Green, Halbert, and Robinson (1965; $\theta_A = .6, .8$) found inferences very close to Bayesian when $N = 1$. The evidence is mixed but with more of the studies leaning toward under-inference.⁵⁸

⁵⁸Presumably, the many papers on base-rate neglect also contain evidence on inferences from samples of size 1. Virtually none of them have 50-50 priors, however, so it is difficult to disentangle biased inference from base-rate neglect. We do not review this literature systematically, but to give a flavor of what it may indicate, we examined Bar-Hillel’s (1980) seminal paper. Our impression is that the evidence in Bar-Hillel’s paper roughly mirrors the evidence from experiments on single-signal inference reviewed above. The full distribution of subjects’ reported posteriors can be eyeballed from histograms reported for Bar-Hillel’s Studies 1, 2, 3, 7 and 8, each of which presents an inference problem where a single signal is indicative of the less likely of two states that subjects are given base rates for. We divide the 222 subjects’ responses into four categories. Because the signal strength always was in the opposite direction of the base rate, the 33% of subjects whose posteriors equaled the base rate or weaker must have been either under-inferring from the signal or (as is presumably unlikely) “over-using” the base rate. By contrast, 9% of subjects reported posteriors stronger than the signal, almost surely indicating over-use of the signal (since otherwise they must be reversing the base rate). 31% of subjects reported posteriors of exactly the signal strength. Although not logically necessary, we share the presumption of Bar-Hillel and most researchers in this area that these subjects were almost surely simply using the signal strength and ignoring the base rate altogether. The remaining 27% of subjects reported posteriors strictly between the correct Bayesian posteriors and the posteriors that would completely ignore the base rate. It is unclear how many of these subjects were over-using or under-using the signal because anyone under-using the base rate could have been either over-inferring or under-inferring from the signal. From these data taken together, it seems likely that between 9% and 36% of the subjects were over-inferring from the signal, and at least 33% of the subjects were under-using the signal.

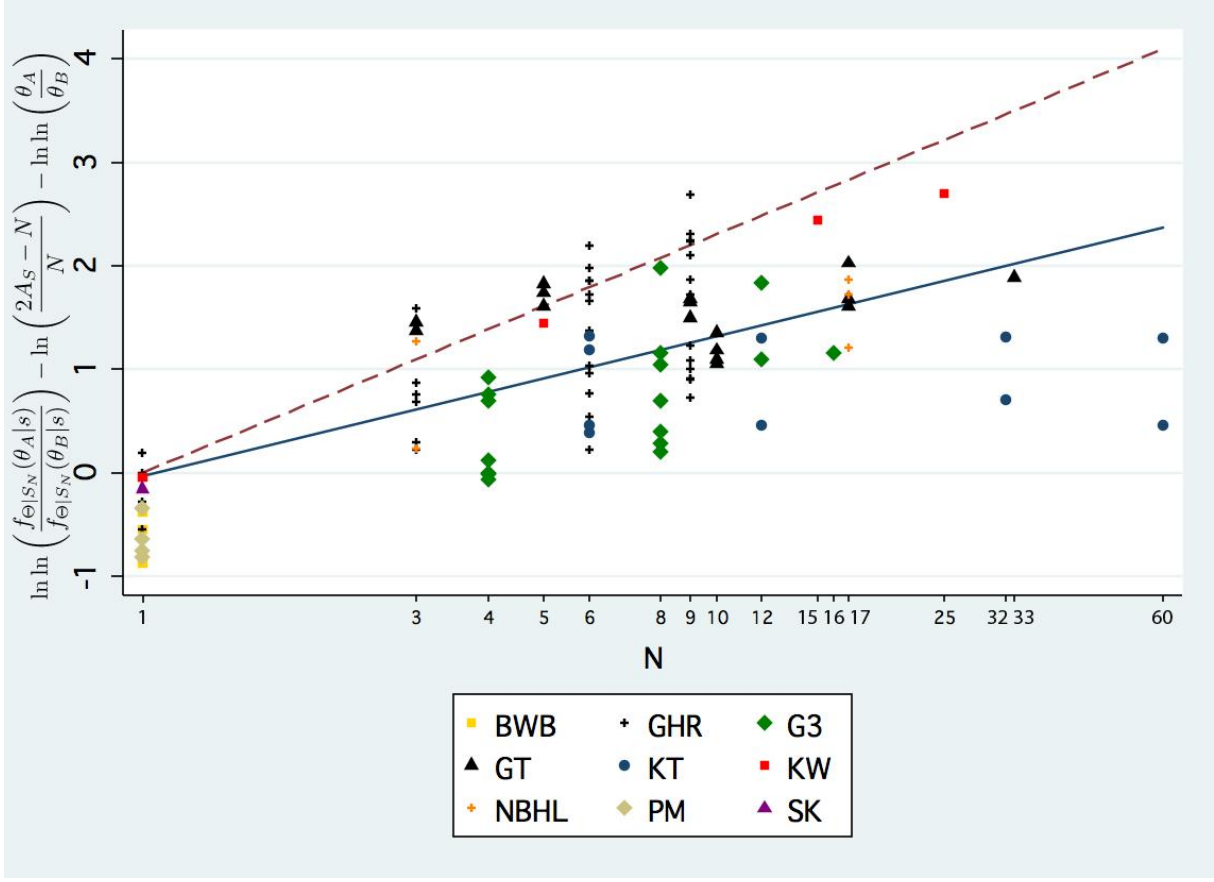


Figure B2: Inference with symmetric rates and equal priors

Notes: The x-axis is depicted on the natural log scale. For all datapoints in the figure, subjects knew that prior probabilities of the two rates were equal. The dotted line represents the null hypothesis of Bayesian updating, and the solid line is the best-fitting regression line from column 2 of Table B2. The includes studies are: BWB = Beach, Wise, & Barclay (1970); GHR = Green, Halbert, and Robinson (1965); G3 = Grether’s (1992) Study 3; GT = Griffin & Tversky’s (1992) Study 1; KT = Kahneman & Tversky (1972); KW = Kraemer & Weber (2004); NBHL = Nelson, Bloomfield, Hales, & Libby’s (2001) Study 1; PM = Peterson & Miller (1965); SK = Sasaki & Kawagoe (2007).

As an aside, we can use the same balls-and-urns experiments to study how priors affect inference and thereby measure the prevalence and extent of base-rate neglect in these studies. Taking the log of Bayes’ Rule and rearranging:

$$\ln \frac{f_{\Theta|S_N}(\theta_A|s)}{f_{\Theta|S_N}(\theta_B|s)} - \ln \frac{f_{S_N|\Theta}(s|\theta_A)}{f_{S_N|\Theta}(s|\theta_B)} = \ln \frac{f_{\Theta}(\theta_A)}{f_{\Theta}(\theta_B)}. \quad (7)$$

As above, from the published experiments, we obtain for each inference problem the values θ_A , θ_B , $f_{\Theta}(\theta_A)$, the actual sample observed by subjects, and subjects' mean or median reported posterior. The right-hand side of equation (7) can be readily calculated from $f_{\Theta}(\theta_A)$, as can the first term on the left-hand side, $\ln \frac{f_{\Theta|S_N}(\theta_A|s)}{f_{\Theta|S_N}(\theta_B|s)}$, from subjects' posterior. In order not to confound base-rate neglect with other biases (that affect inference even when priors are equal), we calculate the second term on the left-hand side, $\ln \frac{f_{S_N|\Theta}(s|\theta_A)}{f_{S_N|\Theta}(s|\theta_B)}$, as the predicted value, $\ln \frac{\widehat{f_{S_N|\Theta}(s|\theta_A)}}{\widehat{f_{S_N|\Theta}(s|\theta_B)}}$, from the previously-estimated regression equation (6). We use only the symmetric-rate data. Figure B3 plots the left-hand side of equation (7) against $\ln \frac{f_{\Theta}(\theta_A)}{f_{\Theta}(\theta_B)}$. Regardless of whatever other biases may affect inference, if the subjects correctly incorporate base rates into their inferences, the data should lie along the identity line (the dashed line in Figure B3). However, the best-fitting line (the solid line in the figure) has a slope less than 1, indicating that subjects' inferences are too insensitive to the prior probabilities.

To formally investigate the degree of base-rate neglect, we rewrite (7) as a regression equation:

$$\ln \frac{f_{\Theta|S_N}(\theta_A|s)}{f_{\Theta|S_N}(\theta_B|s)} - \ln \frac{f_{S_N|\Theta}(s|\theta_A)}{f_{S_N|\Theta}(s|\theta_B)} = \varphi_0 + \varphi_1 \ln \frac{f_{\Theta}(\theta_A)}{f_{\Theta}(\theta_B)} + \varepsilon. \quad (8)$$

The null hypothesis of Bayesian updating corresponds to $\varphi_0 = 0$ and $\varphi_1 = 1$.

The first column of Table B3 estimates equation (8) on the full sample. While we can reject the null hypothesis that $\varphi_0 = 0$, it is clear from Figure B5 that the deviation is small. On the other hand, the fitted value $\widehat{\varphi}_1 = .601$ is significantly less than 1 (and larger than 0). The second column estimates the equation using only the data with unequal prior probabilities, but the results are similar. Finally, the third column restricts the data to incentivized experiments. While $\widehat{\varphi}_0$ is now essentially zero, $\widehat{\varphi}_1 = .405$ indicates even stronger base-rate neglect in these data.

Table B3. Base-rate neglect.

	(1)	(2)	(5)
	All data	Only unequal priors	Only incentivized
$\ln \frac{f_{\Theta}(\theta_A)}{f_{\Theta}(\theta_B)}$	0.524 (0.078)	0.510 (0.081)	0.360 (0.115)
Constant	0.216 (0.053)	0.299 (0.087)	-0.130 (0.084)
R^2	0.242	0.303	0.176
#obs	209	110	48
#papers	10	5	3

Notes: Results are from OLS regressions, with standard errors in parentheses. The dependent variable is as described in the text.

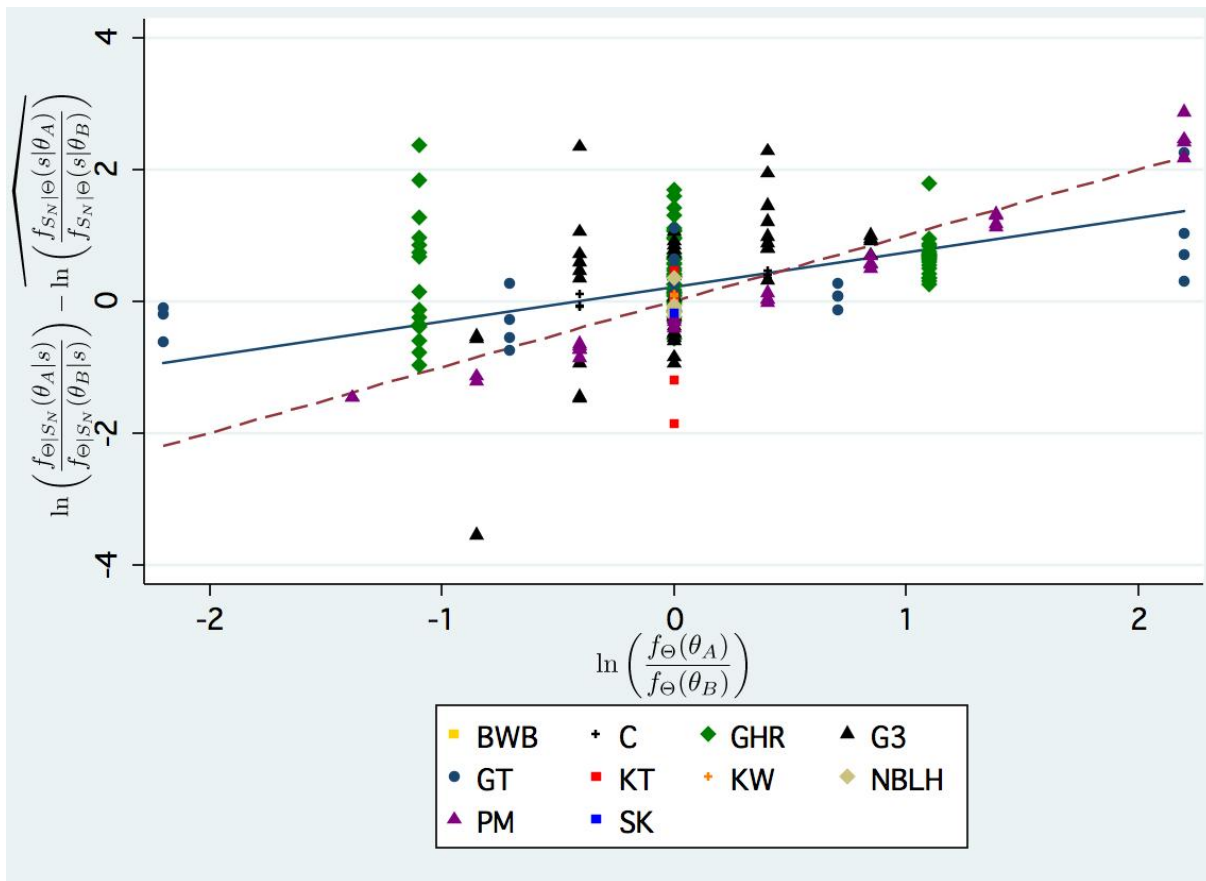


Figure B3: Base-rate neglect in symmetric-rate inference problems.

Notes: The dotted line represents the null hypothesis of Bayesian updating, and the solid line is the best-fitting regression line from column 1 of Table B5. The includes studies are: BWB = Beach,

Wise, & Barclay (1970); C = Camerer (1987); GHR = Green, Halbert, and Robinson (1965); G3 = Grether’s (1992) Study 3; GT = Griffin & Tversky’s (1992) Study 1; KT = Kahneman & Tversky (1972); KW = Kraemer & Weber (2004); NBHL = Nelson, Bloomfield, Hales, & Libby’s (2001) Study 1; PM = Peterson & Miller (1965); SK = Sasaki & Kawagoe (2007).

9.3 Consistency Between Subjective Sampling Distributions and Inference

A core feature of the fully rational model that our model retains is that people draw inferences that are consistent with Bayes’s Rule applied to their subjective sampling distributions. Although surely not a perfect fit, we believe (like other researchers before us) that this feature is approximately right except insofar as people neglect base rates. There is indirect supportive evidence from the qualitative correspondence between the evidence on subjective sampling distributions (reviewed in section B.1) and the evidence on inferences (reviewed in section B.2). There is direct evidence from two studies that measured subjective sampling distributions and inferences for the same subject.

Peterson, DuCharme, & Edwards (1968, Study 2) conducted symmetric inference experiments with every combination of $N = 3, 5, 8$, and $\theta_A = 1 - \theta_B = .6, .7, .8$. Then subjects drew subjective sampling distributions for the nine binomial distributions (shown in Figure B4). Peterson, DuCharme, & Edwards plotted subjects’ inferences against what their inferences would be if they applied Bayes’s Rule to their subjective sampling distributions. Peterson, DuCharme, & Edwards found that “most points cluster extremely close to the identity line.”

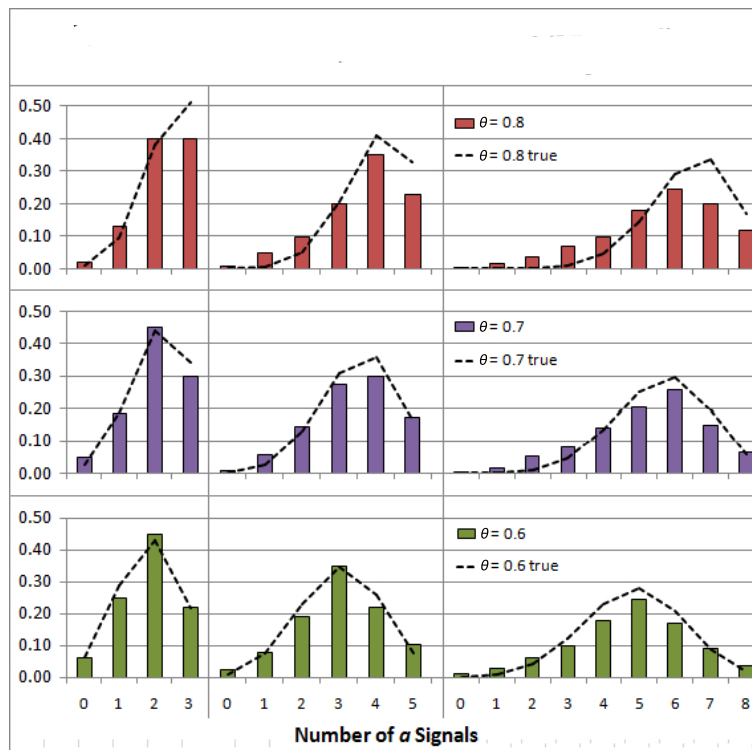


Figure B4: Median probability estimates, $N=3$, $N=5$, and $N=8$ (Peterson DuCharme and Edwards 1968)

Wheeler & Beach (1968) elicited subjects' subjective sampling distributions for a sample of size $N = 8$ for rates $\theta = .6$ and $.8$ (see Figure B5) and then asked subjects to make bets in an inference task. Wheeler & Beach inferred subjective posteriors from the bets, under the assumption that subjects sought to maximize expected winnings. The correlation between the median subjective log-posterior (calculated from the first 20 inference problems) and the median subjective log-likelihood (calculated from applying Bayes's Rule to the subjective sampling distribution elicited at the very beginning of the experiment) was $.90$. For individual subjects' data, the median correlation was $.85$.

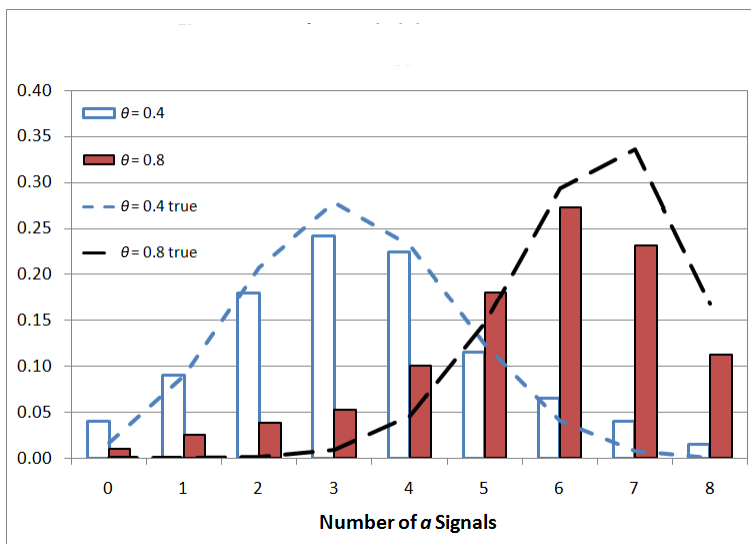


Figure B5: Median probability estimates, $N=8$ (Wheeler and Beach 1968)

9.4 Inference For Sequential Clumps

In Section 4, we lay out various possible dynamic extensions of our model of NBLLN, but there are few experiments aimed at comparing inferences from a sample presented simultaneously with a sample drawn sequentially, and there are no experiments that elicit people's beliefs about what they will infer conditional on observing samples in the future.

Grether (1992, Study 3) confronted subjects with (incentivized) symmetric, binomial inference problems, with rates $\theta_A = 1 - \theta_B = .2, .3, .4, .6, .7, .8$, and priors $f_{\Theta}(\theta_A) = .3, .4, .5, .6, .7$. The sample size always began as $N = 4$. In some cases, however, after subjects made their inference, they were asked to make an updated inference after an additional 4 signals were drawn, up to a maximum of 12 signals in total. Although only aggregate statistics are reported in the paper, David

Grether sent us the subject-level data. In a few cases, we can learn about how subjects process the signals by comparing their inferences before and after they receive a clump. For example, in one situation, the rates were $\theta_A = 1 - \theta_B = .2$, the prior probabilities of the rates were equal, and the first four signals were all b 's. The next four signals were 2 a 's and 2 b 's. The objective posterior probability of rate θ_B is the same after all eight as after the first four: .9961. However, the subjects' subjective posterior (that is, the median across subjects) is .95 after the first four signals and .70 after all eight. This pattern of additional uninformative signals causing the subjective posterior to move toward .5 is consistent with retrospective-pooling Barney, not retrospective-acceptive Barney. The same pattern holds in all three other test cases in Grether's data.⁵⁹

In contrast, Kraemer & Weber (2004) present evidence that supports retrospective-acceptive Barney. For an incentivized, symmetric binomial inference problem with $\theta_A = 1 - \theta_B = .6$ and $N = 5$, subjects presented with a sample of 3 a 's and 2 b 's gave mean posterior probability for rate θ_A of .585. Other subjects who were instead shown the same signals as two separate samples, one with 3 a 's and 0 b 's and one with 0 a 's and 2 b 's, gave mean posterior probability for state A of .56, which is marginally statistically different. Similarly, with $\theta_A = 1 - \theta_B = .6$ and $N = 25$, subjects' mean posterior probability for rate θ_A was .56 when the sample was 13 a 's and 12 b 's, but .53 (strongly statistically distinguishable from .56) when two samples of 13 a 's and 0 b 's, and then 0 a 's and 12 b 's were presented sequentially. The fact that subjects make different inferences in these two cases is inconsistent with being retrospective-pooling, but not inconsistent with being retrospective-acceptive.⁶⁰

Shu & Wu's (2003) Study 3 appears to be inconsistent with any of the dynamic extensions of our model that we consider. They conduct a symmetric binomial inference problem with samples of size $N = 10$ and three different levels of the rates, $\theta_A = 1 - \theta_B = .6, .75, \text{ or } .9$. In one condition,

⁵⁹In the 2nd test case, $f_{\Theta}(\theta_A) = .4$, $\theta_A = 1 - \theta_B = .3$, and the first eight signals were 1 a and 7 b 's. The next four signals were 2 a 's and 2 b 's. While the objective posterior probability of rate θ_B is identically .9908, subjects' median subjective posterior fell from .9 to .5. In the 3rd test case, $f_{\Theta}(\theta_A) = .7$, $\theta_A = 1 - \theta_B = .4$, and the first four signals were 2 a 's and 2 b 's, and the next four signals were 2 a 's and 2 b 's. The objective posterior probability of rate θ_A remains .7, but subjects' median subjective posterior fell from .775 to .7. In the 4th test case, $f_{\Theta}(\theta_A) = .3$, $\theta_A = 1 - \theta_B = .7$, and the first four and next four signals were 2 a 's and 2 b 's. The objective posterior probability of rate θ_A remains .3, but subjects' median subjective posterior increased from .42 to .51. If we examine subjects' mean posterior probabilities, the pattern is robust in the first three test cases (.87 to .66, .72 to .59, and .73 to .68) and hard to interpret in the fourth (.54 to .58).

⁶⁰Unlike the experimental subjects, however, retrospective-acceptive Barney will under-infer weakly *more* from two clumps than from one. This is because an additional clump provides information about an additional draw of β from the distribution, which is weakly more informative about the true rate than an additional signal about a fixed β . That being said, a combination of retrospective-acceptive Barney with SDTSD could explain why the two extreme clumps lead to a weaker overall inference than the single clump.

Kraemer & Weber (2004) also have a third experimental treatment that generates ambiguous evidence. In the $N = 5$ case, when subjects are presented with a sample of 1 a and 1 b , followed by a sample of 2 a 's and 1 b , their mean posterior is .575, in between the other two treatments and not statistically distinguishable from either. Similarly, in the $N = 25$ case, the mean posterior is .555 when subjects are presented with a sample of 6 a 's and 6 b 's, followed by a sample of 7 a 's and 6 b 's.

subjects observed the 10 signals one at a time before stating a posterior belief. In the other conditions, subjects observed the 10 signals in clumps of 2 signals each or 5 signals each. While for some realizations of the 10-signal sets subjects draw less extreme inferences when the signals arrive in larger clumps—as predicted for retrospective-acceptive Barney—the results on average go in the opposite direction.⁶¹ While it may be possible to reconcile Shu & Wu’s results with a combination of NBLLN and the dynamics of base-rate neglect, that combined model should be worked out to see if it systematically reverses some of the conclusions we reach in Section 5.

9.5 Evidence For Non-Binomial Distributions

While the vast majority of simple inference experiments have been conducted with binomial signals, there are a few studies with other distributions. The results overall are consistent with NBLLN and SDTSD applying beyond binomial subjective sampling distributions.

There are a handful of studies where signals are multinomial. For example, in Beach’s (1968) experiment, there were two decks of cards, a Red Deck and a Green Deck. Each card had a letter from A to F written on it. The Red and Green Decks had equal priors, but each deck had different proportions of the lettered cards. The subjects were shown $N = 3$ cards, one card at a time, and reported their subjective probability that the cards were being drawn from the Red Deck, as opposed to the Green Deck, after each draw. The likelihood ratios for the cards ranged from 1:2.5 to 3:1. For example, the likelihood ratio for card F was 1:2. A second group of subjects faced the same inference task with the same likelihood ratios for each card, but with the absolute probabilities scaled down for some cards and scaled up for others. For example, for the first group of subjects, the probability of an F card was .03 for the Red Deck and .06 for the Green Deck; for the second group of subjects, the probability of an F card was .16 for the Red Deck and .32 for the Green Deck. The first main finding is that subjects under-inferred on average. The other finding was that, for a given objective likelihood ratio, Group 1 under-inferred more for cards where Group 1’s probabilities for that card were scaled down relative to Group 2’s. Our interpretation is that when Group 1’s probabilities are scaled down, the observed sample lies further in the tails on the subjective sampling distribution for both decks. SDTSD predicts more extreme under-inference in such cases.

Under-inference is also the general finding in the other multinomial experiments we could find that compared subjects’ posteriors with Bayesian posteriors (Phillips, Hays, & Edwards’s Study 1, 1966; Dale, 1968; Martin, 1969; Martin & Gettys, 1969; Chapman, 1973). However, there are

⁶¹Sanders (1968) and Beach, Wise, & Barclay (1970) also compare inferences from simultaneously-presented samples with inferences from sequentially-presented samples, but it is difficult to interpret their results because the results for the simultaneously-presented samples are averaged across different sample sizes.

two exceptions: (1) Phillips, Hays, & Edwards (1966) varied the sample size of signals observed by subjects and, while finding under-inference for $N = 3, 5,$ and $9,$ found essentially Bayesian inference for $N = 1,$ and (2) Dale (1968) reported that in one particular trial where the data happened to exactly match one of the multinomial rates, about $1/8$ of the subject over-inferred—the most over-inference he observed on any trial. Martin (1969), Martin & Gettys (1969), and Chapman (1973) also reported that subjects’ under-inferred by more when an observed sample warranted a more extreme conclusion. We believe this “discriminability” finding is likely due to SDTSD.

DuCharme (1970) conducted a normal-signal inference experiment. He found under-inference when the sample was relatively far in the tails of both distributions, consistent with SDTSD, although he interpreted his results as meaning that people are reluctant to report extreme probabilities. Gustafson, Shukla, Delbecq, & Walster (1973) told subjects the average heights and weights of Midwestern college-age men and women. Subjects were then asked a series of questions such as, “The observed height of a person is 68 inches. Is the person more likely to be a male or female? How much more likely?” Gustafson et al. found that subjects over-inferred when the objective likelihood ratio was relatively small and under-inferred when the objective likelihood ratio was relatively large. Assuming that subjects believed that the sampling distributions for height and weight were normal distributions, this result means that subjects over-inferred when the sample was relatively close to the men’s or women’s mean height or weight and under-inferred when the sample was relatively far in the tails of both distributions. In two studies, DuCharme & Peterson (1968) familiarized subjects with normal distributions for male and female heights and then elicited subjects’ beliefs that a sample was being drawn from the population of men or of women. Subjects’ posteriors were nearly Bayesian when $N = 1,$ but subjects under-inferred for samples of size $N = 4.$

Peterson & Phillips (1966) conducted an experiment where the rate generating binary signals was drawn from a uniform distribution on $[0, 1].$ Subjects observed 48 binary signals and after each signal had to specify a 33% confidence interval for the rate. Subjects’ confidence intervals were almost always too wide, indicating that subjects under-inferred from the data about the rate.

10 Appendix C: Proofs

10.1 Preliminary and Appendix A Results

According to Whitt's (1979) Theorem 4, A3 implies that for any N and any $\theta_A, \theta_B \in \Theta$, Barney's likelihood ratio $\frac{f_{S_N|\Theta}^\psi(s|\theta_A)}{f_{S_N|\Theta}^\psi(s|\theta_B)}$ is strictly increasing in the number of a signals in the sample. We will repeatedly use this fact in the proofs.

The preliminary results assume that $f_{B|\Theta}^\psi(\beta|\theta)$ has the functional form of the beta distribution with parameters $\theta\psi$ and $(1-\theta)\psi$. Let $p_0 \equiv \frac{f_\Theta(\theta_A)}{f_\Theta(\theta_B)}$ denote the agent's prior ratio, let $\Pi_{S_N|\Theta \times \Theta}^\psi(s|\theta_A, \theta_B) \equiv \frac{f_{S_N|\Theta}^\psi(s|\theta_A)}{f_{S_N|\Theta}^\psi(s|\theta_B)}$ denote Barney's likelihood ratio after observing a clump of N signals $s \in S_N$, and let $\Pi_{\Theta \times \Theta|S_N}^\psi(\theta_A, \theta_B|s) \equiv \frac{f_{\Theta|S_N}^\psi(\theta_A|s)}{f_{\Theta|S_N}^\psi(\theta_B|s)} = p_0 \Pi_{S_N|\Theta \times \Theta}^\psi(s|\theta_A, \theta_B)$ denote his posterior ratio.

Lemma $\beta 1$. *Assume Barney is prospective-acceptive, he believes the rate is θ , and he will observe a clump of N signals $s \in S_N$. Then he believes the probability of observing A_s a -signals in the sequence S_N is:*

$$f_{S_N|\Theta}^\psi(s|\theta_A) = \frac{\Gamma(\psi)}{\Gamma(\psi+N)} \frac{\Gamma(\theta\psi + A_s)}{\Gamma(\theta\psi)} \frac{\Gamma((1-\theta)\psi + N - A_s)}{\Gamma((1-\theta)\psi)} \frac{\Gamma(N+1)}{\Gamma(A_s+1)\Gamma(N-A_s+1)}.$$

Proof:

$$\begin{aligned} & f_{S_N|\Theta}^\psi(s|\theta_A) \\ &= \int_0^1 f_{S_N|B}^\psi(s|\beta) f_{B|\Theta}^\psi(\beta|\theta_A) d\beta \\ &= \int_0^1 \frac{\Gamma(N+1)}{\Gamma(A_s+1)\Gamma(N-A_s+1)} \beta^{A_s} (1-\beta)^{N-A_s} \frac{\Gamma(\psi)}{\Gamma(\theta\psi)\Gamma((1-\theta)\psi)} \beta^{\theta\psi-1} (1-\beta)^{(1-\theta)\psi-1} d\beta \\ &= \frac{\Gamma(\psi)}{\Gamma(\psi+N)} \frac{\Gamma(\theta\psi + A_s)}{\Gamma(\theta\psi)} \frac{\Gamma((1-\theta)\psi + N - A_s)}{\Gamma((1-\theta)\psi)} \frac{\Gamma(N+1)}{\Gamma(A_s+1)\Gamma(N-A_s+1)} \times \\ & \quad \int_0^1 \frac{\Gamma(\psi+N)}{\Gamma(\theta\psi + A_s)\Gamma((1-\theta)\psi + N - A_s)} \beta^{\theta\psi+A_s-1} (1-\beta)^{(1-\theta)\psi+N-A_s-1} d\beta \\ &= \frac{\Gamma(\psi)}{\Gamma(\psi+N)} \frac{\Gamma(\theta\psi + A_s)}{\Gamma(\theta\psi)} \frac{\Gamma((1-\theta)\psi + N - A_s)}{\Gamma((1-\theta)\psi)} \frac{\Gamma(N+1)}{\Gamma(A_s+1)\Gamma(N-A_s+1)}, \end{aligned}$$

where the fourth equality follows because the term being integrated is the pdf of a beta distribution, and the integral of a pdf is equal to 1. □

Lemma $\beta 2$. Consider two states of the world, A and B , associated with known rates θ_A and θ_B , and a prior ratio p_0 . Assume Barney is retrospective pooling. Barney's likelihood ratio after observing a clump of N signals $s \in S_N$ which has A_s a -signals is:

$$\Pi_{S_N|\Theta \times \Theta}^\psi (s|\theta_A, \theta_B) = \frac{\Gamma(\theta_A \psi + A_s) \Gamma((1 - \theta_A) \psi + N - A_s) \Gamma(\theta_B \psi) \Gamma((1 - \theta_B) \psi)}{\Gamma(\theta_B \psi + A_s) \Gamma((1 - \theta_B) \psi + N - A_s) \Gamma(\theta_A \psi) \Gamma((1 - \theta_A) \psi)}.$$

Moreover, in the "symmetric inference" case where $\theta_A = 1 - \theta_B$,

$$\Pi_{S_N|\Theta \times \Theta}^\psi (s|\theta_A, \theta_B) = \frac{\Gamma(\theta_A \psi + A_s) \Gamma((1 - \theta_A) \psi + N - A_s)}{\Gamma(\theta_B \psi + A_s) \Gamma((1 - \theta_B) \psi + N - A_s)}$$

Proof: This is an immediate implication of the previous lemma. □

Lemma $\beta 3$. Consider two states of the world, A and B , associated with known rates $\theta_A \geq \theta_B$. Suppose Barney has observed N signals, s_N , and he observes his $N + 1^{\text{st}}$ signal. Then Barney's likelihood ratio is closer to $\Pi_{S_N|\Theta \times \Theta}^\psi (s_N|\theta_A, \theta_B)$ if he pools the $N + 1^{\text{st}}$ signal with the others than if he groups it separately: that is, if the $N + 1^{\text{st}}$ signal is an a -signal, then

$$\Pi_{S_1|\Theta \times \Theta}^\psi (a|\theta_A, \theta_B) \Pi_{S_N|\Theta \times \Theta}^\psi (s_N|\theta_A, \theta_B) \geq \Pi_{S_{N+1}|\Theta \times \Theta}^\psi (s_N \cup a|\theta_A, \theta_B),$$

and if $N + 1^{\text{st}}$ signal is a b -signal, then

$$\Pi_{S_1|\Theta \times \Theta}^\psi (b|\theta_A, \theta_B) \Pi_{S_N|\Theta \times \Theta}^\psi (s_N|\theta_A, \theta_B) \leq \Pi_{S_{N+1}|\Theta \times \Theta}^\psi (s_N \cup a|\theta_A, \theta_B).$$

Proof: We will prove the claim in the lemma that pertains if the final signal is an a -signal; the proof for the other case is analogous. Denote by A_N the number of a -signals in s_N . If Barney pools

the a -signal with the others, then his likelihood ratio is

$$\begin{aligned}
& \Pi_{S_{N+1}|\Theta \times \Theta}^\psi (s_N \cup a | \theta_A, \theta_B) \\
= & \frac{\Gamma(\theta_A \psi + A_N + 1) \Gamma((1 - \theta_A) \psi + N + 1 - A_N - 1) \Gamma(\theta_B \psi) \Gamma((1 - \theta_B) \psi)}{\Gamma(\theta_B \psi + A_N + 1) \Gamma((1 - \theta_B) \psi + N + 1 - A_N - 1) \Gamma(\theta_A \psi) \Gamma((1 - \theta_A) \psi)} \\
= & \frac{(\theta_A \psi + A_N) \Gamma(\theta_A \psi + A_N) \Gamma((1 - \theta_A) \psi + N - A_N) \Gamma(\theta_B \psi) \Gamma((1 - \theta_B) \psi)}{(\theta_B \psi + A_N) \Gamma(\theta_B \psi + A_N) \Gamma((1 - \theta_B) \psi + N - A_N) \Gamma(\theta_A \psi) \Gamma((1 - \theta_A) \psi)} \\
= & \frac{\theta_A \psi + A_N}{\theta_B \psi + A_N} \Pi_{S_N|\Theta \times \Theta}^\psi (s_N | \theta_A, \theta_B),
\end{aligned}$$

where the second equality uses Lemma $\beta 2$, and the third equality uses the fact that $\Gamma(n+1) = n\Gamma(n)$.

If Barney instead separately groups the final signal, then his likelihood ratio is

$$\begin{aligned}
\Pi_{S_1|\Theta \times \Theta}^\psi (a | \theta_A, \theta_B) \Pi_{S_N|\Theta \times \Theta}^\psi (s_N | \theta_A, \theta_B) &= \frac{\theta_A \psi}{\theta_B \psi} \Pi_{S_N|\Theta \times \Theta}^\psi (s_N | \theta_A, \theta_B) \\
&= \frac{\theta_A}{\theta_B} \Pi_{S_N|\Theta \times \Theta}^\psi (s_N | \theta_A, \theta_B).
\end{aligned}$$

By assumption, $\frac{\theta_A}{\theta_B} \geq 1$. Therefore, for all $k > 0$,

$$\frac{\theta_A}{\theta_B} \geq \frac{\theta_A \psi + k}{\theta_B \psi + k} \geq 1.$$

The result follows. □

Lemma $\beta 4$. *Consider two states of the world, A and B , associated with known rates for $\theta_A > \theta_B$, and a prior ratio p_0 . Assume Barney is retrospective pooling.*

- *Barney draws the same inferences as Tommy when the observed sequence is ab or ba .*
- *If the sample is all a 's or all b 's, Barney under-infers relative to Tommy.*
- *For any true state $\theta \in (0, 1)$, Barney believes that: as $N \rightarrow \infty$, his posterior ratio $\frac{f_{\Theta|S_N}^\psi(\theta_A|s)}{f_{\Theta|S_N}^\psi(\theta_B|s)}$ will converge in distribution to a random variable that has full support on $(0, \infty)$.*

Proof:

- Barney's posterior after one a and one b signal is

$$\begin{aligned}
& p_0 \frac{\Gamma(\theta_A \psi + 1) \Gamma((1 - \theta_A) \psi + 1) \Gamma(\theta_B \psi) \Gamma((1 - \theta_B) \psi)}{\Gamma(\theta_B \psi + 1) \Gamma((1 - \theta_B) \psi + 1) \Gamma(\theta_A \psi) \Gamma((1 - \theta_A) \psi)} \\
&= p_0 \frac{\Gamma(\theta_A \psi) \theta_A \psi \Gamma((1 - \theta_A) \psi) (1 - \theta_A) \psi \Gamma(\theta_B \psi) \Gamma((1 - \theta_B) \psi)}{\Gamma(\theta_B \psi) \theta_B \psi \Gamma((1 - \theta_B) \psi) (1 - \theta_B) \psi \Gamma(\theta_A \psi) \Gamma((1 - \theta_A) \psi)} \\
&= p_0 \frac{\theta_A (1 - \theta_A)}{\theta_B (1 - \theta_B)},
\end{aligned}$$

which equals Tommy's posterior.

- Barney's posterior after N a -signals and 0 b -signals is:

$$\begin{aligned}
& p_0 \frac{\Gamma(\theta_A \psi + N) \Gamma((1 - \theta_A) \psi) \Gamma(\theta_B \psi) \Gamma((1 - \theta_B) \psi)}{\Gamma(\theta_B \psi + N) \Gamma((1 - \theta_B) \psi) \Gamma(\theta_A \psi) \Gamma((1 - \theta_A) \psi)} \\
&= p_0 \frac{\Gamma(\theta_A \psi + N) \Gamma(\theta_B \psi)}{\Gamma(\theta_B \psi + N) \Gamma(\theta_A \psi)} \\
&= p_0 \frac{(\theta_A \psi + N - 1)(\theta_A \psi + N - 2) \dots (\theta_A \psi) \Gamma(\theta_A \psi) \Gamma(\theta_B \psi)}{(\theta_B \psi + N - 1)(\theta_B \psi + N - 2) \dots (\theta_B \psi) \Gamma(\theta_B \psi) \Gamma(\theta_A \psi)} \\
&= p_0 \frac{(\theta_A \psi + N - 1)(\theta_A \psi + N - 2) \dots (\theta_A \psi)}{(\theta_B \psi + N - 1)(\theta_B \psi + N - 2) \dots (\theta_B \psi)}.
\end{aligned}$$

Note that the numerator and denominator each has N terms in it. Tommy's posterior ratio is

$$p_0 \frac{(\theta_A)^N}{(\theta_B)^N},$$

so also for Tommy, the numerator and denominator each has N terms. Furthermore, since $\frac{\theta_A}{\theta_B} > 1$, for all $k > 0$, $\frac{\theta_A + k}{\theta_B + k} < \frac{\theta_A}{\theta_B}$. Therefore,

$$p_0 \frac{(\theta_A \psi + N - 1)(\theta_A \psi + N - 2) \dots (\theta_A \psi)}{(\theta_B \psi + N - 1)(\theta_B \psi + N - 2) \dots (\theta_B \psi)} < p_0 \frac{(\theta_A)^N}{(\theta_B)^N}.$$

Moreover, the likelihood ratio clearly favors state A for both Tommy and Barney. Hence, Barney under-infers relative to Tommy. The case of N b -signals and 0 a -signals proceeds analogously.

- Lemma 1 implies that Barney's subjective sampling distribution for an infinite sample has pdf $\lim_{N \rightarrow \infty} f_{S_N | \Theta}^\psi \left(\frac{A_s}{N} = \alpha | \theta \right) = f_{B | \Theta}^\psi (\beta = \alpha | \theta)$ a.s. Therefore, after observing proportion α of a -signals in a large sample, Barney anticipates having likelihood ratio $\frac{f_{B | \Theta}^\psi (\beta = \alpha | \theta_A)}{f_{B | \Theta}^\psi (\beta = \alpha | \theta_B)}$ a.s. By

A3, $\frac{f_{\mathbb{B}|\Theta}^\psi(\beta=\alpha|\theta_A)}{f_{\mathbb{B}|\Theta}^\psi(\beta=\alpha|\theta_B)}$ is strictly increasing in α .

Furthermore,

$$\begin{aligned} \frac{f_{\mathbb{B}|\Theta}^\psi(\beta=0|\theta_A)}{f_{\mathbb{B}|\Theta}^\psi(\beta=0|\theta_B)} &= \frac{\frac{\Gamma(\psi)}{\Gamma(\theta_A\psi)\Gamma((1-\theta_A)\psi)} 0^{\theta_A\psi-1} (1-0)^{(1-\theta_A)\psi-1}}{\frac{\Gamma(\psi)}{\Gamma(\theta_B\psi)\Gamma((1-\theta_B)\psi)} 0^{\theta_B\psi-1} (1-0)^{(1-\theta_B)\psi-1}} \\ &= \frac{\Gamma(\theta_B\psi)\Gamma((1-\theta_B)\psi) 0^{(\theta_A-\theta_B)\psi}}{\Gamma(\theta_A\psi)\Gamma((1-\theta_A)\psi) 1^{(\theta_A-\theta_B)\psi}} = 0 \end{aligned}$$

and

$$\begin{aligned} \frac{f_{\mathbb{B}|\Theta}^\psi(\beta=1|\theta_A)}{f_{\mathbb{B}|\Theta}^\psi(\beta=1|\theta_B)} &= \frac{\frac{\Gamma(\psi)}{\Gamma(\theta_A\psi)\Gamma((1-\theta_A)\psi)} 1^{\theta_A\psi-1} 0^{(1-\theta_A)\psi-1}}{\frac{\Gamma(\psi)}{\Gamma(\theta_B\psi)\Gamma((1-\theta_B)\psi)} 1^{\theta_B\psi-1} 0^{(1-\theta_B)\psi-1}} \\ &= \frac{\Gamma(\theta_B\psi)\Gamma((1-\theta_B)\psi) 1^{(\theta_A-\theta_B)\psi}}{\Gamma(\theta_A\psi)\Gamma((1-\theta_A)\psi) 0^{(\theta_A-\theta_B)\psi}} = \infty. \end{aligned}$$

According to A1, Barney's beliefs put full support on $\alpha \in (0, 1)$. Hence Barney thinks his large-sample likelihood ratio will converge in distribution to a random variable whose support is $\left(\frac{f_{\mathbb{B}|\Theta}^\psi(\beta=0|\theta_A)}{f_{\mathbb{B}|\Theta}^\psi(\beta=0|\theta_B)}, \frac{f_{\mathbb{B}|\Theta}^\psi(\beta=1|\theta_A)}{f_{\mathbb{B}|\Theta}^\psi(\beta=1|\theta_B)}\right) = (0, \infty)$. Since Barney's posterior ratio is his likelihood ratio times his prior ratio, the result follows. □

Lemma $\beta 5$. *Parameterized-Barney, whose subjective-rate distribution has the beta-distribution functional form*

$$f_{\mathbb{B}|\Theta}^\psi(\beta|\theta) = \frac{\Gamma(\psi)}{\Gamma(\theta\psi)\Gamma((1-\theta)\psi)} \beta^{\theta\psi-1} (1-\beta)^{(1-\theta)\psi-1},$$

satisfies A1-A4.

Proof: We will prove each property in turn.

- Clearly the full-support property holds since $f_{\mathbb{B}|\Theta}^\psi(\beta|\theta) = \frac{\Gamma(\psi)}{\Gamma(\theta\psi)\Gamma((1-\theta)\psi)} \beta^{\theta\psi-1} (1-\beta)^{(1-\theta)\psi-1} > 0$ for all $\beta \in (0, 1)$ and $\theta \in (0, 1)$. This function is also clearly point-wise continuous in θ since all the components are point-wise continuous. Since $f_{\mathbb{B}|\Theta}^\psi(\beta|\theta)$ is constructed to be a pdf, it is Lebesgue-integrable with respect to β and hence absolutely continuous with respect to β .
- Note that

$$f_{\mathbb{B}|\Theta}^\psi(\beta|\theta) = \frac{\Gamma(\psi)}{\Gamma(\theta\psi)\Gamma((1-\theta)\psi)} \beta^{\theta\psi-1} (1-\beta)^{(1-\theta)\psi-1} = f_{\mathbb{B}|\Theta}^\psi(1-\beta|1-\theta).$$

Therefore,

$$\int_0^{1-x} f_{\mathbb{B}|\Theta}^\psi(\beta|\theta) d\beta = \int_0^{1-x} f_{\mathbb{B}|\Theta}^\psi(1-\beta|1-\theta) d(1-\beta).$$

The left-hand side is $F_{\mathbb{B}|\Theta}^\psi(1-x|\theta)$. Using the change-of-variables $\tilde{\beta} = 1-\beta$, the right-hand side becomes

$$\begin{aligned} -\int_1^x f_{\mathbb{B}|\Theta}^\psi(\tilde{\beta}|1-\theta) d\tilde{\beta} &= \int_x^1 f_{\mathbb{B}|\Theta}^\psi(\tilde{\beta}|1-\theta) d\tilde{\beta} \\ &= 1 - \int_0^x f_{\mathbb{B}|\Theta}^\psi(\tilde{\beta}|1-\theta) d\tilde{\beta}. \end{aligned}$$

Hence

$$F_{\mathbb{B}|\Theta}^\psi(1-x|\theta) = 1 - F_{\mathbb{B}|\Theta}^\psi(x|1-\theta).$$

- Note that

$$\begin{aligned} \frac{f_{\mathbb{B}|\Theta}^\psi(\beta|\theta')}{f_{\mathbb{B}|\Theta}^\psi(\beta|\theta)} &= \frac{\frac{\Gamma(\psi)}{\Gamma(\theta'\psi)\Gamma((1-\theta')\psi)}\beta^{\theta'\psi-1}(1-\beta)^{(1-\theta')\psi-1}}{\frac{\Gamma(\psi)}{\Gamma(\theta\psi)\Gamma((1-\theta)\psi)}\beta^{\theta\psi-1}(1-\beta)^{(1-\theta)\psi-1}} \\ &= \frac{\Gamma(\theta\psi)\Gamma((1-\theta)\psi)}{\Gamma(\theta'\psi)\Gamma((1-\theta')\psi)} \left(\frac{\beta}{1-\beta}\right)^{(\theta'-\theta)\psi}. \end{aligned}$$

If $\theta' > \theta$, then this expression is increasing in β .

- The mean of the beta distribution with parameters $\theta\psi$ and $(1-\theta)\psi$ is $\frac{\theta\psi}{\theta\psi+(1-\theta)\psi} = \theta$.

□

Lemma A. Fix $\theta_A \geq \theta_B$ and some $0 < \underline{\lambda} < \bar{\lambda} < \infty$. If Barney's posterior ratio after one signal is in $(\underline{\lambda}, \bar{\lambda})$,

$$[p_0\Pi_{S_1|\Theta \times \Theta}^\psi(b|\theta_A, \theta_B), p_0\Pi_{S_1|\Theta \times \Theta}^\psi(a|\theta_A, \theta_B)] \subseteq (\underline{\lambda}, \bar{\lambda}),$$

then for any natural number N , we can construct a sequence of N signals, $s \in S_N$, such that for every truncation of s —i.e., for every sequence comprising only the first $n \leq N$ signals of s —the posterior ratio after that truncation, $\Pi_{\Theta \times \Theta|S_n}^\psi(\theta_A, \theta_B|s_n)$, is in $(\underline{\lambda}, \bar{\lambda})$.

Proof: We prove this by induction. By hypothesis, the result is true for the truncation $n = 1$. Assume that the likelihood ratio after N signals is in $(\underline{\lambda}, \bar{\lambda})$, and we will show that we can add a signal, and it will still be in $(\underline{\lambda}, \bar{\lambda})$. Suppose that there are A_N a -signals and $N - A_N$ b -signals. Since

$$[p_0\Pi_{S_1|\Theta \times \Theta}^\psi(b|\theta_A, \theta_B), p_0\Pi_{S_1|\Theta \times \Theta}^\psi(a|\theta_A, \theta_B)] \subseteq (\underline{\lambda}, \bar{\lambda})$$

and $\Pi_{S_1|\Theta \times \Theta}^\psi(b|\theta_A, \theta_B) \leq 1 \leq \Pi_{S_1|\Theta \times \Theta}^\psi(a|\theta_A, \theta_B)$, it follows that $p_0 \in (\underline{\lambda}, \bar{\lambda})$.

There are two cases. The first is that Barney's posterior ratio after the sequence of N signals is (weakly) less than p_0 , and so $\Pi_{S_N|\Theta \times \Theta}^\psi(s_N|\theta_A, \theta_B) \leq 1$. In this case add an a -signal. The new posterior ratio is $p_0 \Pi_{S_{N+1}|\Theta \times \Theta}^\psi(s_N \cup a|\theta_A, \theta_B)$. Now,

$$\begin{aligned} \bar{\lambda} \geq p_0 \Pi_{S_1|\Theta \times \Theta}^\psi(a|\theta_A, \theta_B) &\geq p_0 \Pi_{S_1|\Theta \times \Theta}^\psi(a|\theta_A, \theta_B) \Pi_{S_N|\Theta \times \Theta}^\psi(s_N|\theta_A, \theta_B) \\ &\geq p_0 \Pi_{S_{N+1}|\Theta \times \Theta}^\psi(s_N \cup a|\theta_A, \theta_B) \\ &\geq p_0 \Pi_{S_N|\Theta \times \Theta}^\psi(s_N|\theta_A, \theta_B) \geq \underline{\lambda} \end{aligned}$$

The first inequality is by hypothesis, the second is due to the fact that $\Pi_{S_N|\Theta \times \Theta}^\psi(s_N|\theta_A, \theta_B) \leq 1$, the third inequality is due to Lemma $\beta 3$, the fourth inequality holds since the posterior ratio increases when an additional signal is an a -signal, and the last inequality is again by hypothesis.

The second case is that the likelihood ratio after N signals is greater than p_0 , in which case add a b -signal, and the argument is analogous. □

Lemma B. Fix $\theta_A \geq \theta_B$. If $p_0 \in \left[\frac{1}{\Pi_{S_1|\Theta \times \Theta}^\psi(a|\theta_A, \theta_B)}, \frac{1}{\Pi_{S_1|\Theta \times \Theta}^\psi(b|\theta_A, \theta_B)} \right]$, then for any N , there exists a sequence of signals s_i for $i = 0, 1, \dots, N$, where s_i is a truncation of s_N , such that for all $i = 0, 1, \dots, N$, $p_0 \Pi_{S_i|\Theta \times \Theta}^\psi(s_i|\theta_A, \theta_B) \in \left[\frac{1}{\Pi_{S_1|\Theta \times \Theta}^\psi(a|\theta_A, \theta_B)}, \frac{1}{\Pi_{S_1|\Theta \times \Theta}^\psi(b|\theta_A, \theta_B)} \right]$.

Proof: The lemma is a direct implication of the following claim: If

$$p_0 \in \left[\frac{1}{\Pi_{S_1|\Theta \times \Theta}^\psi(a|\theta_A, \theta_B)}, \frac{1}{\Pi_{S_1|\Theta \times \Theta}^\psi(b|\theta_A, \theta_B)} \right],$$

then for any n , there exists a sequence of signals $\{s_i\}_{i=0,1,\dots,n}$, such that

$$p_0 \Pi_{S_{n+1}|\Theta \times \Theta}^\psi(s_n \cup a|\theta_A, \theta_B) \geq 1$$

and

$$p_0 \Pi_{S_{n+1}|\Theta \times \Theta}^\psi(s_n \cup b|\theta_A, \theta_B) \leq 1.$$

In words, we can construct a sequence of signals such that Barney's posterior ratio flips around 1 with each signal. We will prove this claim by induction and then show that it implies the lemma.

Clearly for $n = 0$ the claim is true, since the prior ratio is in $\left[\frac{1}{\Pi_{S_1|\Theta \times \Theta}^\psi(a|\theta_A, \theta_B)}, \frac{1}{\Pi_{S_1|\Theta \times \Theta}^\psi(b|\theta_A, \theta_B)} \right]$. Now we will assume the statement is true up to n and prove it is true for $n + 1$.

There are two cases. In the first case, assume that after n signals, it is the a -signal that

flips the posterior ratio around 1: $p_0 \Pi_{S_n | \Theta \times \Theta}^\psi (s_n | \theta_A, \theta_B) \leq 1$ and $p_0 \Pi_{S_{n+1} | \Theta \times \Theta}^\psi (s_n \cup a | \theta_A, \theta_B) \geq 1$ (since adding an a -signal must increase the posterior ratio). Assume that the claim is not true for $n + 1$: that is, there is no set of $n + 1$ signals for which $p_0 \Pi_{S_{n+2} | \Theta \times \Theta}^\psi (s_{n+1} \cup a | \theta_A, \theta_B) \geq 1$ and $p_0 \Pi_{S_{n+2} | \Theta \times \Theta}^\psi (s_{n+1} \cup b | \theta_A, \theta_B) \leq 1$. In particular, taking the set of n signals plus an a -signal,

$$p_0 \Pi_{S_{n+2} | \Theta \times \Theta}^\psi (s_n \cup a \cup a | \theta_A, \theta_B) \geq 1$$

and

$$p_0 \Pi_{S_{n+2} | \Theta \times \Theta}^\psi (s_n \cup a \cup b | \theta_A, \theta_B) \geq 1.$$

Because the claim is true up to n signals, however, we know that an additional b -signal must flip the posterior ratio below 1,

$$p_0 \Pi_{S_{n+1} | \Theta \times \Theta}^\psi (s_n \cup b | \theta_A, \theta_B) \leq 1.$$

But then we have identified a set of $n + 1$ signals for which the statement is true, namely $s_n \cup b$: we know that $s_n \cup b \cup b$ must also generate a posterior ratio below 1 (since adding another b -signal must decrease the likelihood ratio); and $s_n \cup b \cup a$ generates the same posterior ratio as $s_n \cup a \cup b$, which we know is above 1. So we have a contradiction. Therefore, either $s_n \cup a$ or $s_n \cup b$ must satisfy the statement. The proof for a b -signal proceeds analogously.

To see that the statement implies the lemma, assume WLOG that $p_0 \Pi_{S_n | \Theta \times \Theta}^\psi (s_n | \theta_A, \theta_B) \leq 1$, and that $p_0 \Pi_{S_{n+1} | \Theta \times \Theta}^\psi (s_n \cup a | \theta_A, \theta_B) \geq 1$. Since

$$p_0 \Pi_{S_{n+1} | \Theta \times \Theta}^\psi (s_n \cup a | \theta_A, \theta_B) \leq p_0 \Pi_{S_n | \Theta \times \Theta}^\psi (s_n | \theta_A, \theta_B) \Pi_{S_1 | \Theta \times \Theta}^\psi (a | \theta_A, \theta_B)$$

(due to Lemma $\beta 3$), it follows that

$$1 \leq p_0 \Pi_{S_n | \Theta \times \Theta}^\psi (s_n | \theta_A, \theta_B) \Pi_{S_1 | \Theta \times \Theta}^\psi (a | \theta_A, \theta_B).$$

Hence

$$\frac{1}{\Pi_{S_1 | \Theta \times \Theta}^\psi (a | \theta_A, \theta_B)} \leq p_0 \Pi_{S_n | \Theta \times \Theta}^\psi (s_n | \theta_A, \theta_B) \leq 1 \leq \frac{1}{\Pi_{S_1 | \Theta \times \Theta}^\psi (b | \theta_A, \theta_B)}.$$

The proof for the case that a b -signal flips the posterior ratio below 1 proceeds analogously. \square

Lemma C. Fix rates $\theta_A \geq \theta_B$, prior ratio p_0 , and the true state $i \in \{A, B\}$ (with corresponding true rate θ_i). Denote Barney's limit likelihood ratio conditional on the true state as $\Pi_{S_\infty | \Theta \times \Theta}^\psi (s_\infty | \theta_A, \theta_B, \Omega = i)$. Suppose Barney is retrospective-pooling. For any $0 < \underline{\lambda} < \bar{\lambda} < \infty$

satisfying the following two statements:

1. $p_0 \Pi_{S_\infty | \Theta \times \Theta}^\psi (s_\infty | \theta_A, \theta_B, \Omega = i) \in (\underline{\lambda}, \bar{\lambda})$, and
2. $[p_0 \Pi_{S_1 | \Theta \times \Theta}^\psi (b | \theta_A, \theta_B), p_0 \Pi_{S_1 | \Theta \times \Theta}^\psi (a | \theta_A, \theta_B)] \subseteq (\underline{\lambda}, \bar{\lambda})$ or $p_0 \in \left[\frac{1}{\Pi_{S_1 | \Theta \times \Theta}^\psi (a | \theta_A, \theta_B)}, \frac{1}{\Pi_{S_1 | \Theta \times \Theta}^\psi (b | \theta_A, \theta_B)} \right] \subseteq (\underline{\lambda}, \bar{\lambda})$,

there is a positive probability that for all $N \in \{1, 2, \dots\}$,

$$\Pi_{\Theta \times \Theta | S_N}^\psi (\theta_A, \theta_B | s_N) \in (\underline{\lambda}, \bar{\lambda}).$$

Proof: The proof proceeds in three steps:

1. Show that the prior ratio is in $(\underline{\lambda}, \bar{\lambda})$. This is true under either of the assumptions in condition 2 of the lemma.

- Since $\Pi_{S_1 | \Theta \times \Theta}^\psi (b | \theta_A, \theta_B) \leq 1$ and $\Pi_{S_1 | \Theta \times \Theta}^\psi (a | \theta_A, \theta_B) \geq 1$, the assumption

$$[p_0 \Pi_{S_1 | \Theta \times \Theta}^\psi (b | \theta_A, \theta_B), p_0 \Pi_{S_1 | \Theta \times \Theta}^\psi (a | \theta_A, \theta_B)] \subseteq (\underline{\lambda}, \bar{\lambda})$$

implies $p_0 \in (\underline{\lambda}, \bar{\lambda})$.

- Alternatively, we directly assume that $p_0 \in \left[\frac{1}{\Pi_{S_1 | \Theta \times \Theta}^\psi (a | \theta_A, \theta_B)}, \frac{1}{\Pi_{S_1 | \Theta \times \Theta}^\psi (b | \theta_A, \theta_B)} \right] \subseteq (\underline{\lambda}, \bar{\lambda})$.
- Note: If $\theta_A = \theta_B$, then Barney's posterior equals his prior for any sequence of signals. Since $p_0 \in (\underline{\lambda}, \bar{\lambda})$, the conclusion of the lemma follows trivially. Hence, for the remainder of the proof, assume $\theta_A > \theta_B$.

2. Show that for any finite N , we can construct a sequence of signals such that the posterior ratio is always in $(\underline{\lambda}, \bar{\lambda})$. There are two ways we can guarantee this:

- If $[p_0 \Pi_{S_1 | \Theta \times \Theta}^\psi (b | \theta_A, \theta_B), p_0 \Pi_{S_1 | \Theta \times \Theta}^\psi (a | \theta_A, \theta_B)] \subseteq (\underline{\lambda}, \bar{\lambda})$, then it is true by Lemma A.
- Alternatively, if we assume $p_0 \in \left[\frac{1}{\Pi_{S_1 | \Theta \times \Theta}^\psi (a | \theta_A, \theta_B)}, \frac{1}{\Pi_{S_1 | \Theta \times \Theta}^\psi (b | \theta_A, \theta_B)} \right] \subseteq (\underline{\lambda}, \bar{\lambda})$, then it is true by Lemma B.

3. Show that there exists a large enough number of signals \hat{N} such that if the posterior ratio after \hat{N} signals is in $(\underline{\lambda}, \bar{\lambda})$, then with positive probability the posterior ratio will always be inside $(\underline{\lambda}, \bar{\lambda})$. The rest of the proof demonstrates this. The conclusion of the lemma then

follows from combining steps 1-3, taking the sequence in step 2 to have \hat{N} signals, which (because it is a finite sequence) has positive probability.

Let $g(r) \equiv \Pi_{S_\infty|\Theta \times \Theta}^\psi(s_\infty|\theta_A, \theta_B, r)$ denote Barney's likelihood ratio in an infinite sample where the proportion of a -signals is r . This function will allow us to map from an observed proportion r to the posterior ratio that it would imply for an infinite sample, $p_0g(r)$. Even in dealing with finite samples, it will be useful to work with the $g(\cdot)$ function in describing the mapping from the proportion of a -signals to Barney's posterior ratio. In a finite sample, however, $p_0g(r)$ will only approximate Barney's posterior ratio, but we can bound the approximation error in a large enough sample; formally, using the Law of Large Numbers, for any r , $\epsilon > 0$, and $\nu > 0$, there exists an $N_{r,\epsilon,\nu}$ such that if $N \geq N_{r,\epsilon,\nu}$, then with probability at least $1 - \nu$, the posterior ratio after observing a proportion r a -signals out of N is within ϵ of $p_0g(r)$. Since all we will need for step 3 is that the posterior ratio remains within $(\underline{\lambda}, \bar{\lambda})$ with positive probability, the ν we pick does not matter. So fix ν .

Since $p_0g(r)$ is only an approximation of Barney's posterior ratio for a finite sample, we need to impose tighter bounds than $(\underline{\lambda}, \bar{\lambda})$ on where $p_0g(r)$ can wander in a finite sample in order to ensure that Barney's posterior ratio always remains within $(\underline{\lambda}, \bar{\lambda})$. To that end, fixing ϵ , we will now pick two other numbers, $\underline{\tau} < \bar{\tau}$, such that:

- $\underline{\tau} > \underline{\lambda} + \epsilon$, and $\bar{\tau} < \bar{\lambda} - \epsilon$;
- the limit posterior ratio conditional on the true state, $p_0\Pi_{S_\infty|\Theta \times \Theta}^\psi(s_\infty|\theta_A, \theta_B, \Omega = i)$, is in $(\underline{\tau}, \bar{\tau})$; and
- $g^{-1}(\underline{\tau})$ and $g^{-1}(\bar{\tau})$ are both rational.

Note that $\underline{\tau}$ and $\bar{\tau}$ satisfying the above properties exist because g is continuous and monotonic in r .

We now map these bounds on the posterior ratio onto the implied bounds on the proportion of a -signals. Let \underline{r} denote the proportion of a -signals such that $p_0g(\underline{r}) = \underline{\tau}$; and \bar{r} , the proportion of a -signals such that $p_0g(\bar{r}) = \bar{\tau}$. By assumption both \underline{r} and \bar{r} are rational. Denote the proportion of realized a -signals after N signals by r_N .

We claim that for any particular \underline{r} and \bar{r} such that the true rate $\theta_i \in (\underline{r}, \bar{r})$, there is a large enough number of signals \tilde{N} such that if there have already been \tilde{N} observed with a proportion a -signals $r_{\tilde{N}} \in (\underline{r}, \bar{r})$, then with positive probability, $r_n \in (\underline{r}, \bar{r})$ for all numbers of signals $n \geq \tilde{N}$. This is because by Chebyshev's inequality, for any N and any $\delta > 0$, $P(|r_N - \theta_i| < \delta) \geq 1 - \frac{\theta_i(1-\theta_i)}{N^2\delta^2}$.

Thus, the probability that the proportion of a -signals always stays within the bounds \underline{r}, \bar{r} is

$$P\left(|r_n - \theta_i| < \delta \forall n \geq \tilde{N} \text{ s.t. } |r_{\tilde{N}} - \theta_i| < \delta\right) \geq \prod_{n=\tilde{N}}^{\infty} \left(1 - \frac{\theta_i(1 - \theta_i)}{n^2\delta^2}\right).$$

Note that for any $x > 0$ and $y < 1$, it is true that $y(1 - x) > y - x$, and so we can bound the right-hand side of the previous equation by iteratively applying this inequality:

$$\prod_{n=\tilde{N}}^{\infty} \left(1 - \frac{\theta_i(1 - \theta_i)}{n^2\delta^2}\right) > 1 - \sum_{n=\tilde{N}}^{\infty} \frac{\theta_i(1 - \theta_i)}{n^2\delta^2}.$$

Hence there is positive probability that the proportion of a -signals always stays within (\underline{r}, \bar{r}) . While not relevant for this proof, we point out for reference in proofs below that this infinite series is a known convergent sum; so as $\tilde{N} \rightarrow \infty$, the right-hand side converges to 1.

Therefore, for any ϵ , we can choose a large enough number of signals, $\hat{N} = \max\{\tilde{N}, N_{\underline{r}, \epsilon, \nu}, N_{\bar{r}, \epsilon, \nu}\}$, so that if the proportion of a -signals after \hat{N} signals is within (\underline{r}, \bar{r}) , then with positive probability the proportion always remains within (\underline{r}, \bar{r}) . Therefore, with positive probability $p_0g(r)$ always remains within $(\underline{\tau}, \bar{\tau})$. Since $\hat{N} \geq N_{\underline{r}, \epsilon, \nu}$ and $\hat{N} \geq N_{\bar{r}, \epsilon, \nu}$, whenever $p_0g(r)$ is within $(\underline{\tau}, \bar{\tau})$, Barney's posterior ratio is within $(\underline{\lambda}, \bar{\lambda})$. □

Lemma A1. *Barney-Freddy does not believe in LLN: For any interval $[\alpha_1, \alpha_2] \subseteq [0, 1]$,*

$$\lim_{N \rightarrow \infty} \sum_{x=\lfloor a_1 N \rfloor}^{\lfloor a_2 N \rfloor} f_{S_N | \Theta}^{\psi M}(A_s = x | \theta) = F_{\beta | \Theta}^{\psi}(\beta = \alpha_2 | \theta) - F_{\beta | \Theta}^{\psi}(\beta = \alpha_1 | \theta) > 0.$$

Proof: First, we claim that conditional on a particular subjective-rate β , the agent believes that the distribution of percentage of a -signals converges to a point mass on β . To see this, fix β , and suppose Barney-Freddy has observed N signals. We begin with the case where N is an even number (i.e., the agent has observed $\frac{N}{2}$ pairs of signals, where recall that the ‘‘urn’’ is renewed after each pair of signals). Considering each pair of signals as a single signal that has 4 possible values and a mean value of 2β , LLN implies that Barney-Freddy believes that the distribution of percentage of a -signals converges to a point mass on β . This proves the claim for the case where N is an even number. Now, notice that for any $[\alpha_1, \alpha_2] \subseteq [0, 1]$ and $\epsilon > 0$, we can find a large enough \hat{N} such that for all $N \geq \hat{N}$, if $\frac{A_{sN}}{N} \in (\alpha_1, \alpha_2)$, then $\frac{A_{sN}}{N+1} \in (\alpha_1 - \epsilon, \alpha_2 + \epsilon)$ and $\frac{A_{sN+1}}{N+1} \in (\alpha_1 - \epsilon, \alpha_2 + \epsilon)$. This observation, combined with the claim for even number N , proves the claim for odd number N .

The remainder of the proof is identical to the proof of Lemma 1 below. □

10.2 Main Text Results

Lemma 1. *Barney does not believe in LLN: For any interval $[\alpha_1, \alpha_2] \subseteq [0, 1]$,*

$$\lim_{N \rightarrow \infty} \sum_{x=\lfloor \alpha_1 N \rfloor}^{\lceil \alpha_2 N \rceil} f_{S_N | \Theta}^\psi (A_s = x | \theta) = F_{\mathbb{B} | \Theta}^\psi (\beta = \alpha_2 | \theta) - F_{\mathbb{B} | \Theta}^\psi (\beta = \alpha_1 | \theta) > 0.$$

Proof: We begin by proving a key fact. Consider some rational $\alpha \in [0, 1]$ (which will represent the percentage of a -signals) and an increasing sequence of integers m_1, m_2, \dots (which will represent sample sizes) such that αm_j (which will represent numbers of a -signals) is an integer for all $j = 1, 2, \dots$. Using the definition of Barney-ness,

$$\begin{aligned} \lim_{j \rightarrow \infty} F_{S_{m_j} | \Theta}^\psi (\alpha m_j | \theta) &= \lim_{j \rightarrow \infty} \int_0^1 F_{S_{m_j} | \mathbb{B}} (\alpha m_j | \beta) f_{\mathbb{B} | \Theta}^\psi (\beta | \theta) d\beta \\ &= \int_0^1 \lim_{j \rightarrow \infty} F_{S_{m_j} | \mathbb{B}} (\alpha m_j | \beta) f_{\mathbb{B} | \Theta}^\psi (\beta | \theta) d\beta \\ &= \int_0^\alpha \lim_{j \rightarrow \infty} F_{S_{m_j} | \mathbb{B}} (\alpha m_j | \beta) f_{\mathbb{B} | \Theta}^\psi (\beta | \theta) d\beta + \int_\alpha^1 \lim_{j \rightarrow \infty} F_{S_{m_j} | \mathbb{B}} (\alpha m_j | \beta) f_{\mathbb{B} | \Theta}^\psi (\beta | \theta) d\beta \\ &= \int_0^\alpha (1) f_{\mathbb{B} | \Theta}^\psi (\beta | \theta) d\beta + \int_\alpha^1 (0) f_{\mathbb{B} | \Theta}^\psi (\beta | \theta) d\beta \\ &= F_{\mathbb{B} | \Theta}^\psi (\beta = \alpha | \theta), \end{aligned}$$

where the second equality follows from the absolute continuity of $f_{\mathbb{B} | \Theta}^\psi (\beta | \theta)$, and the fourth follows from the Law of Large Numbers.

Now, by the above fact and the definition of the cdf, for any $[a_1, a_2] \subseteq [0, 1]$ with a_1, a_2 rational,

$$\begin{aligned} \lim_{N \rightarrow \infty} \sum_{x=\lfloor a_1 N \rfloor}^{\lceil a_2 N \rceil} f_{S_N | \Theta}^\psi (A_s = x | \theta) &= \lim_{N \rightarrow \infty} (F_{S_N | \Theta}^\psi (\lceil a_2 N \rceil | \theta) - F_{S_N | \Theta}^\psi (\lfloor a_1 N \rfloor | \theta)) \\ &= F_{\mathbb{B} | \Theta}^\psi (\beta = a_2 | \theta) - F_{\mathbb{B} | \Theta}^\psi (\beta = a_1 | \theta). \end{aligned}$$

Consider a sequence of pairs of rational numbers, $(a_1^{(1)}, a_2^{(1)})$, $(a_1^{(2)}, a_2^{(2)})$, ... that converges to the

pair of real numbers (α_1, α_2) . Taking the limit of the above equality along this sequence gives

$$\lim_{N \rightarrow \infty} \sum_{x=\lfloor \alpha_1 N \rfloor}^{\lceil \alpha_2 N \rceil} f_{S_N|\Theta}^\psi(A_s = x|\theta) = F_{\beta|\Theta}^\psi(\beta = \alpha_2|\theta) - F_{\beta|\Theta}^\psi(\beta = \alpha_1|\theta).$$

This is greater than 0 by the full-support assumption in A1. □

Proposition 1. *For any $\theta \in (0, 1)$ and $N \in \{1, 2, \dots\}$:*

1. $E_{S_N|\Theta}^\psi\left(\frac{A_s}{N}|\theta\right) = E_{S_N|\Theta}\left(\frac{A_s}{N}|\theta\right) = \theta$.
2. $F_{S_N|\Theta}^\psi(A_s|\theta)$ SOSD $F_{S_N|\Theta}(A_s|\theta)$, and $\text{Var}_{S_N|\Theta}^\psi\left(\frac{A_s}{N}|\theta\right) \geq \text{Var}_{S_N|\Theta}\left(\frac{A_s}{N}|\theta\right)$ with strict inequality for $N > 1$.
3. $\text{Var}_{S_N|\Theta}^\psi\left(\frac{A_s}{N}|\theta\right)$ is strictly decreasing in N , and $\lim_{N \rightarrow \infty} \text{Var}_{S_N|\Theta}^\psi\left(\frac{A_s}{N}|\theta\right) > 0$.
4. $F_{S_N|\Theta}^\psi(A_s|\theta')$ FOSD $F_{S_N|\Theta}^\psi(A_s|\theta)$ whenever $\theta' > \theta$.

Proof:

1. Since A_s is a binomial random variable with rate θ , $E_{S_N|\Theta}\left(\frac{A_s}{N}|\theta\right) = \theta$. Using the Law of Iterated Expectations, $E_{S_N|\Theta}^\psi\left(\frac{A_s}{N}|\theta\right) = E_{\beta|\Theta}^\psi[E_{S_N|\beta}\left(\frac{A_s}{N}|\beta\right)|\theta] = E_{\beta|\Theta}^\psi[\beta|\theta] = \theta$, where the last equality follows from A4.
2. To keep notation compact, let T denote a random variable whose distribution is binomial distribution with rate θ . Let Y denote the random variable whose distribution is $F_{S_N|\Theta}^\psi(A_s|\theta)$, i.e., the random variable induced by taking a binomial draw using rate β and then integrating over all possible β 's using $f_{\beta|\Theta}^\psi(\beta|\theta)$. We will show that Y is a mean preserving spread of the binomial distribution with rate θ , which implies second-order stochastic dominance.

For Y to be a mean preserving spread of T , it must be the case the $Y = T + Z$, where Z is a random variable with conditional mean $E[Z|T] = 0$. We will construct Z .

Recall that for any random variable V , its moment generating function $M_V(t) \equiv E[e^{tV}]$ (when it exists) completely characterizes the distribution of V and has the following useful properties: $E[V] = \frac{d}{dt} M_V(t)|_{t=0}$; for random variables V, V' and V'' , $V = V' + V''$ if and only if $M_V(t) = M_{V'}(t)M_{V''}(t)$; and for random variables V and V' , $M_V(t) = E[M_{V|V'}(t)|V']$.

Since the random variable $Y|\beta$ has a binomial distribution with parameter β , its moment generating function is $M_{Y|\beta}(t) = 1 - \beta + \beta e^t$. Therefore,

$$M_Y(t) = E [M_{Y|\beta}(t)|\beta] = \int_0^1 (1 - \beta + \beta e^t) f_{\mathbb{B}|\Theta}^\psi(\beta|\theta) d\beta,$$

which clearly exists since $M_{Y|\beta}(t)$ exists for each $\beta \in [0, 1]$. If there is some Z , then we know it must have moment generating function

$$M_Z(t) = \frac{M_Y(t)}{M_T(t)} = \frac{\int_0^1 (1 - \beta + \beta e^t) f_{\mathbb{B}|\Theta}^\psi(\beta|\theta) d\beta}{1 - \theta + \theta e^t}.$$

(With this moment generating function, the random variable $T + Z$ will have the same distribution as Y .) Hence, we will have proved that $F_{S_N|\Theta}^\psi(A_s|\theta)$ SOSD $F_{S_N|\Theta}(A_s|\theta)$ once we verify that $E[Z|T] = 0$. By construction, the random variable Z is independent of T (since Z 's moment generating function, and hence its distribution, does not depend on the realization of T). Therefore, $E[Z|T] = E[Z]$. Finally,

$$\begin{aligned} E[Z] &= \left. \frac{d}{dt} M_Z(t) \right|_{t=0} \\ &= \left. \frac{(1 - \theta + \theta e^t) \int_0^1 (\beta e^t) f_{\mathbb{B}|\Theta}^\psi(\beta|\theta) d\beta - \theta e^t \int_0^1 (1 - \beta + \beta e^t) f_{\mathbb{B}|\Theta}^\psi(\beta|\theta) d\beta}{(1 - \theta + \theta e^t)^2} \right|_{t=0} \\ &= \frac{\int_0^1 \beta f_{\mathbb{B}|\Theta}^\psi(\beta|\theta) d\beta - \theta \int_0^1 f_{\mathbb{B}|\Theta}^\psi(\beta|\theta) d\beta}{(1 - \theta + \theta)^2} = \theta - \theta = 0. \end{aligned}$$

This proves the first claim.

Using the Law of Total Variance,

$$\begin{aligned} \text{Var}_{S_N|\Theta}^\psi\left(\frac{A_s}{N}|\theta\right) &= E_{\mathbb{B}|\Theta}^\psi\left[\text{Var}_{S_N|\mathbb{B}}\left(\frac{A_s}{N}|\beta\right)|\theta\right] + \text{Var}_{\mathbb{B}|\Theta}^\psi\left[E_{S_N|\mathbb{B}}\left(\frac{A_s}{N}|\beta\right)|\theta\right] \\ &= E_{\mathbb{B}|\Theta}^\psi\left[\frac{\beta(1-\beta)}{N}|\theta\right] + \text{Var}_{\mathbb{B}|\Theta}^\psi[\beta|\theta] \\ &= \frac{\theta}{N} - \frac{\text{Var}_{\mathbb{B}|\Theta}^\psi[\beta|\theta] + \theta^2}{N} + \text{Var}_{\mathbb{B}|\Theta}^\psi[\beta|\theta] \\ &= \frac{\theta(1-\theta)}{N} + \frac{N-1}{N} \text{Var}_{\mathbb{B}|\Theta}^\psi[\beta|\theta], \end{aligned}$$

where the third equality uses $E_{\mathbb{B}|\Theta}^\psi(\beta^2|\theta) = \text{Var}_{\mathbb{B}|\Theta}^\psi[\beta|\theta] + [E_{\mathbb{B}|\Theta}^\psi(\beta|\theta)]^2$. Since $\text{Var}_{S_N|\Theta}\left(\frac{A_s}{N}|\theta\right) = \frac{\theta(1-\theta)}{N}$, the result immediately follows.

3. From part 2,

$$\text{Var}_{S_N|\Theta}^{\psi} \left(\frac{A_s}{N} | \theta \right) = \frac{\theta(1-\theta)}{N} + \frac{N-1}{N} \text{Var}_{B|\Theta}^{\psi} [\beta | \theta].$$

Note that the highest variance of any distribution that has support on $[0, 1]$ and a mean of θ can have is $\theta(1-\theta)$; this is the variance of a distribution with mass $(1-\theta)$ on 0 and mass θ on 1. Any shifting of weight that preserves the mean must move positive mass closer to the mean. Since $f_{B|\Theta}^{\psi}$ has full support (by A1) and mean θ (by A4), $\text{Var}_{B|\Theta}^{\psi} [\beta | \theta] < \theta(1-\theta)$. Also, the full support of $f_{B|\Theta}^{\psi}$ implies that $\text{Var}_{B|\Theta}^{\psi} [\beta | \theta] > 0$. The result immediately follows.

4.

$$\begin{aligned} F_{S_N|\Theta}^{\psi} (A_s | \theta) &= \sum_{i=0}^{A_s} \int_0^1 \frac{N!}{(N-i)!i!} \beta^i (1-\beta)^{N-i} f_{B|\Theta}^{\psi} (\beta | \theta) d\beta \\ &= \int_0^1 \left[\sum_{i=0}^{A_s} \frac{N!}{(N-i)!i!} \beta^i (1-\beta)^{N-i} \right] f_{B|\Theta}^{\psi} (\beta | \theta) d\beta \\ &= \int_0^1 F_{S_N|B} (A_s | \beta) f_{B|\Theta}^{\psi} (\beta | \theta) d\beta, \end{aligned}$$

where $F_{S_N|B} (A_s | \beta)$ is the cdf of the binomial distribution conditional on a rate β . By A3, $F_{B|\Theta}^{\psi} (\beta | \theta')$ FOSD $F_{B|\Theta}^{\psi} (\beta | \theta)$ whenever $\theta' > \theta$, and therefore $\int_0^1 g(\beta) f_{B|\Theta}^{\psi} (\beta | \theta') d\beta > \int_0^1 g(\beta) f_{B|\Theta}^{\psi} (\beta | \theta) d\beta$ for any function $g(\beta)$ that is decreasing in β . Since $F_{S_N|B} (A_s | \beta)$ is decreasing in β , the result follows. □

Proposition 2. *Let $\theta \in (0, 1)$ be the true rate. Then for any $\theta_A, \theta_B \in \Theta$, Barney draws limited inference even from an infinite sample: as $N \rightarrow \infty$,*

$$\frac{f_{\Theta|S_N}^{\psi} (\theta_A | s)}{f_{\Theta|S_N}^{\psi} (\theta_B | s)} \xrightarrow{\text{a.s.}} \frac{f_{B|\Theta}^{\psi} (\beta = \theta | \theta_A) f(\theta_A)}{f_{B|\Theta}^{\psi} (\beta = \theta | \theta_B) f(\theta_B)},$$

which is a positive, finite number.

Proof: Using the Law of Large Numbers,

$$\lim_{N \rightarrow \infty} \frac{f_{\Theta|S_N}^{\psi} (\theta_A | s)}{f_{\Theta|S_N}^{\psi} (\theta_B | s)} = \left(\lim_{N \rightarrow \infty} \frac{f_{S_N|\Theta}^{\psi} (s | \theta_A)}{f_{S_N|\Theta}^{\psi} (s | \theta_B)} \right) \frac{f(\theta_A)}{f(\theta_B)} = \frac{f_{B|\Theta}^{\psi} (\beta = \theta | \theta_A) f(\theta_A)}{f_{B|\Theta}^{\psi} (\beta = \theta | \theta_B) f(\theta_B)} \text{ a.s.}$$

□

Proposition 3. Fix any $\theta_A > \theta_B \in \Theta$. Barney believes that, regardless of what the true rate $\theta \in (0, 1)$ is, as $N \rightarrow \infty$, his posterior ratio $\frac{f_{\Theta|S_N}^\psi(\theta_A|s)}{f_{\Theta|S_N}^\psi(\theta_B|s)}$ will converge in distribution to a random variable whose support is a simply connected subset of $[0, \infty]$. Moreover, he believes that this subset of $[0, \infty]$ has strictly positive measure.

Proof: Lemma 1 implies that Barney's subjective sampling distribution for an infinite sample is $\lim_{N \rightarrow \infty} f_{S_N|\Theta}^\psi(\frac{A_s}{N} = \alpha|\theta) = f_{\mathbb{B}|\Theta}^\psi(\beta = \alpha|\theta)$ a.s. Therefore, after observing proportion α of a -signals in a large sample, Barney anticipates having posterior ratio $\lim_{N \rightarrow \infty} \frac{f_{\Theta|S_N}^\psi(\theta_A|\frac{A_s}{N}=\alpha)}{f_{\Theta|S_N}^\psi(\theta_B|\frac{A_s}{N}=\alpha)} = \frac{f_{\mathbb{B}|\Theta}^\psi(\beta=\alpha|\theta_A)f(\theta_A)}{f_{\mathbb{B}|\Theta}^\psi(\beta=\alpha|\theta_B)f(\theta_B)}$ a.s. This large-sample posterior ratio is a random variable because, according to A1, Barney's beliefs put full support on $\alpha \in (0, 1)$.

By A3 and Whitt (1979), $\frac{f_{\mathbb{B}|\Theta}^\psi(\beta=\alpha|\theta_A)f(\theta_A)}{f_{\mathbb{B}|\Theta}^\psi(\beta=\alpha|\theta_B)f(\theta_B)}$ is strictly increasing in α . It follows that

$$\inf_{\alpha \in (0,1)} \frac{f_{\mathbb{B}|\Theta}^\psi(\beta = \alpha|\theta_A) f(\theta_A)}{f_{\mathbb{B}|\Theta}^\psi(\beta = \alpha|\theta_B) f(\theta_B)} = \lim_{\alpha \rightarrow 0} \frac{f_{\mathbb{B}|\Theta}^\psi(\beta = \alpha|\theta_A) f(\theta_A)}{f_{\mathbb{B}|\Theta}^\psi(\beta = \alpha|\theta_B) f(\theta_B)} < 1$$

$$\sup_{\alpha \in (0,1)} \frac{f_{\mathbb{B}|\Theta}^\psi(\beta = \alpha|\theta_A) f(\theta_A)}{f_{\mathbb{B}|\Theta}^\psi(\beta = \alpha|\theta_B) f(\theta_B)} = \lim_{\alpha \rightarrow 1} \frac{f_{\mathbb{B}|\Theta}^\psi(\beta = \alpha|\theta_A) f(\theta_A)}{f_{\mathbb{B}|\Theta}^\psi(\beta = \alpha|\theta_B) f(\theta_B)} > 1.$$

Furthermore, by A1 and the Intermediate Value Theorem, the posterior ratio must take on each value in

$$\left(\inf_{\alpha \in (0,1)} \frac{f_{\mathbb{B}|\Theta}^\psi(\beta = \alpha|\theta_A) f(\theta_A)}{f_{\mathbb{B}|\Theta}^\psi(\beta = \alpha|\theta_B) f(\theta_B)}, \sup_{\alpha \in (0,1)} \frac{f_{\mathbb{B}|\Theta}^\psi(\beta = \alpha|\theta_A) f(\theta_A)}{f_{\mathbb{B}|\Theta}^\psi(\beta = \alpha|\theta_B) f(\theta_B)} \right)$$

for some $\alpha \in (0, 1)$. This interval is a simply connected subset of $[0, \infty]$ that Barney believes has strictly positive measure. □

Proposition 4. Fix rates $\theta_A > \theta_B$. For $N = 1$, Barney and Tommy infer the same. If $\theta_A = 1 - \theta_B$, then for any set of $N \in \{1, 2, \dots\}$ signal realizations $s \in S_N$, neither Tommy's beliefs nor Barney's beliefs change from the priors when $\frac{A_s}{N} = \frac{1}{2}$.

Proof: The first claim can be seen from the fact that the for a single sample, the subjective sampling distribution of Tommy matches that of Barney.

To see that the second claim is true, suppose that Barney observes exactly half a and b signals,

k of each signal (so that $N - k = k$). Barney's likelihood ratio is

$$\begin{aligned}
\frac{\int_0^1 \frac{N!}{k!k!} \beta^k (1-\beta)^k f_{\mathbb{B}|\Theta}^\psi(\beta|\theta_A) d\beta}{\int_0^1 \frac{N!}{k!k!} \beta^k (1-\beta)^k f_{\mathbb{B}|\Theta}^\psi(\beta|\theta_B) d\beta} &= \frac{\int_0^1 \frac{N!}{k!k!} \beta^k (1-\beta)^k f_{\mathbb{B}|\Theta}^\psi(\beta|\theta_A) d\beta}{\int_0^1 \frac{N!}{k!k!} \beta^k (1-\beta)^k f_{\mathbb{B}|\Theta}^\psi(\beta|1-\theta_A) d\beta} \\
&= \frac{\int_0^1 \frac{N!}{k!k!} \beta^k (1-\beta)^k f_{\mathbb{B}|\Theta}^\psi(\beta|\theta_A) d\beta}{\int_0^1 \frac{N!}{k!k!} \beta^k (1-\beta)^k f_{\mathbb{B}|\Theta}^\psi(1-\beta|\theta_A) d\beta} \\
&= \frac{\int_0^1 \frac{N!}{k!k!} \beta^k (1-\beta)^k f_{\mathbb{B}|\Theta}^\psi(\beta|\theta_A) d\beta}{\int_0^1 \frac{N!}{k!k!} (1-\beta)^k \beta^k f_{\mathbb{B}|\Theta}^\psi(\beta|\theta_A) d\beta} \\
&= 1,
\end{aligned}$$

where the first equality comes from substituting $\theta_B = 1 - \theta_A$, and the second equality follows from A2. This directly implies that Barney's likelihood ratio is equal to 1 if and only if Tommy's is, since the likelihood ratio is strictly increasing in the number of a -signals for both Barney and Tommy if $\theta_A > \frac{1}{2}$, and equal to 1 for any sample for both Barney and Tommy if $\theta_A = \frac{1}{2}$. \square

Proposition 5. *Fix rates $\theta_A > \theta_B$ and a set of $N \in \{1, 2, \dots\}$ signal realizations $s \in S_N$. If $f_{\mathbb{B}|\Theta}^\psi(\beta|\theta)$ has the functional form (3), then regardless of whether the true rate is θ_A or θ_B , for ψ sufficiently small, the expected change in Barney's beliefs is smaller than the expected change in Tommy's beliefs. Furthermore, suppose $\theta_A = 1 - \theta_B$. Then for any sample of $N > 1$ signals such that $\frac{A_s}{N} \neq \frac{1}{2}$ and any ψ , Barney under-infers relative to Tommy. In addition, while Tommy's inference depends solely on the difference in the number of a and b signals, Barney's change in beliefs is smaller from larger samples with the same difference.*

Proof: We will prove the first claim in two steps. First, we shall show that a sufficient condition is that Tommy's subjective sampling distribution is Blackwell-sufficient for Barney's. Second, we will show that this sufficient condition in fact holds for ψ sufficiently small.

First, suppose $F_{S_N|\Theta}(A_s|\theta_i)$ is Blackwell-sufficient for $F_{S_N|\Theta}^\psi(A_s|\theta_i)$ for $i \in \{A, B\}$. Using Blackwell (1951, 1953), note that $F_{S_N|\Theta}(A_s|\theta_i)$ is Blackwell-sufficient for $F_{S_N|\Theta}^\psi(A_s|\theta_i)$ if and only if for any continuous, convex function g ,

$$\begin{aligned}
&\sum_{k=0}^N f_{S_N|\Theta}(A_s = k|\theta_i) g\left(\frac{f_{S_N|\Theta}(A_s = k|\theta_i)}{f_{S_N|\Theta}(A_s = k|\theta_{-i})}\right) \\
&\geq \sum_{k=0}^N f_{S_N|\Theta}^\psi(A_s = k|\theta_i) g\left(\frac{f_{S_N|\Theta}^\psi(A_s = k|\theta_i)}{f_{S_N|\Theta}^\psi(A_s = k|\theta_{-i})}\right)
\end{aligned}$$

for $i \in \{A, B\}$ and $-i$ being the other state. Taking $g(x) = |p_0 - p_0x|$, this implies

$$\begin{aligned} & \sum_{k=0}^N f_{S_N|\Theta}(A_s = k|\theta_i) \left| p_0 - p_0 \frac{f_{S_N|\Theta}(A_s = k|\theta_i)}{f_{S_N|\Theta}(A_s = k|\theta_{-i})} \right| \\ & \geq \sum_{k=0}^N f_{S_N|\Theta}^\psi(A_s = k|\theta_{-i}) \left| p_0 - p_0 \frac{f_{S_N|\Theta}^\psi(A_s = k|\theta_i)}{f_{S_N|\Theta}^\psi(A_s = k|\theta_{-i})} \right| \end{aligned}$$

for $i \in \{A, B\}$.

From Proposition 1, we know that $f_{S_N|\Theta}^\psi(A_s|\theta_{-i})$ is a mean-preserving spread of $f_{S_N|\Theta}(A_s|\theta_{-i})$.

Since $g(x) = |p_0 - p_0x|$ is a continuous convex function, it follows that

$$\begin{aligned} & \sum_{k=0}^N f_{S_N|\Theta}^\psi(A_s = k|\theta_{-i}) \left| p_0 - p_0 \frac{f_{S_N|\Theta}(A_s = k|\theta_i)}{f_{S_N|\Theta}(A_s = k|\theta_{-i})} \right| \\ & \geq \sum_{k=0}^N f_{S_N|\Theta}(A_s = k|\theta_{-i}) \left| p_0 - p_0 \frac{f_{S_N|\Theta}^\psi(A_s = k|\theta_i)}{f_{S_N|\Theta}^\psi(A_s = k|\theta_{-i})} \right| \end{aligned}$$

for $i \in \{A, B\}$. Chaining the two inequalities,

$$\begin{aligned} & \sum_{k=0}^N f_{S_N|\Theta}(A_s = k|\theta_i) \left| p_0 - p_0 \frac{f_{S_N|\Theta}(A_s = k|\theta_i)}{f_{S_N|\Theta}(A_s = k|\theta_{-i})} \right| \\ & \geq \sum_{k=0}^N f_{S_N|\Theta}(A_s = k|\theta_{-i}) \left| p_0 - p_0 \frac{f_{S_N|\Theta}^\psi(A_s = k|\theta_i)}{f_{S_N|\Theta}^\psi(A_s = k|\theta_{-i})} \right|. \end{aligned}$$

Aggregating over states weighted by the priors,

$$\begin{aligned} & f(\theta_A) \left(\sum_{k=0}^N f_{S_N|\Theta}(A_s = k|\theta_A) \left| p_0 - p_0 \frac{f_{S_N|\Theta}(A_s = k|\theta_A)}{f_{S_N|\Theta}(A_s = k|\theta_B)} \right| \right) \\ & + f(\theta_B) \left(\sum_{k=0}^N f_{S_N|\Theta}(A_s = k|\theta_B) \left| p_0 - p_0 \frac{f_{S_N|\Theta}(A_s = k|\theta_B)}{f_{S_N|\Theta}(A_s = k|\theta_A)} \right| \right) \\ & \geq f(\theta_A) \left(\sum_{k=0}^N f_{S_N|\Theta}(A_s = k|\theta_A) \left| p_0 - p_0 \frac{f_{S_N|\Theta}^\psi(A_s = k|\theta_A)}{f_{S_N|\Theta}^\psi(A_s = k|\theta_B)} \right| \right) \\ & + f(\theta_B) \left(\sum_{k=0}^N f_{S_N|\Theta}(A_s = k|\theta_B) \left| p_0 - p_0 \frac{f_{S_N|\Theta}^\psi(A_s = k|\theta_B)}{f_{S_N|\Theta}^\psi(A_s = k|\theta_A)} \right| \right). \end{aligned}$$

This is what we were seeking to prove.

Now we will show that for $i \in \{A, B\}$ and $N > 1$, $F_{S_N|\Theta}(A_s|\theta_i)$ is Blackwell-sufficient for

$F_{S_N|\Theta}^\psi(A_s|\theta_i)$ if ψ is sufficiently small. Since Barney's subjective sampling distribution is the same as Tommy's for $N = 1$, $F_{S_1|\Theta}(A_s|\theta_i)$ is Blackwell-equivalent to $F_{S_1|\Theta}^\psi(A_s|\theta_i)$ (regardless of the value of ψ). Note that as $\psi \rightarrow 0$, Barney's subjective sampling distribution converges to point masses at 0 and 1. Hence in the limit $\psi \rightarrow 0$, for any N , Barney expects to observe all a 's or all b 's; therefore, in the limit $\psi \rightarrow 0$, $F_{S_N|\Theta}^\psi(A_s|\theta_i)$ is Blackwell-equivalent to $F_{S_1|\Theta}^\psi(A_s|\theta_i)$. In contrast, for Tommy, $F_{S_N|\Theta}(A_s|\theta_i)$ for $N > 1$ is strictly Blackwell-sufficient for $F_{S_1|\Theta}(A_s|\theta_i)$ (e.g., Shaked and Tong, 1990). Therefore, for any $N > 1$, in the limit $\psi \rightarrow 0$, $F_{S_N|\Theta}(A_s|\theta_i)$ is Blackwell-sufficient for $F_{S_N|\Theta}^\psi(A_s|\theta_i)$. Using the necessary and sufficient inequality for Blackwell-sufficiency: for any $N > 1$ and any continuous, convex function g , in the limit $\psi \rightarrow 0$,

$$\begin{aligned} & \sum_{k=0}^N f_{S_N|\Theta}(A_s = k|\theta_i) g\left(\frac{f_{S_N|\Theta}(A_s = k|\theta_i)}{f_{S_N|\Theta}(A_s = k|\theta_{-i})}\right) \\ & > \sum_{k=0}^N f_{S_N|\Theta}^\psi(A_s = k|\theta_i) g\left(\frac{f_{S_N|\Theta}^\psi(A_s = k|\theta_i)}{f_{S_N|\Theta}^\psi(A_s = k|\theta_{-i})}\right). \end{aligned}$$

Note that the right-hand side of this inequality is continuous in ψ . Therefore, the strict inequality holds for any sufficiently small $\psi > 0$, and so $F_{S_N|\Theta}(A_s|\theta_i)$ is Blackwell-sufficient for $F_{S_N|\Theta}^\psi(A_s|\theta_i)$ for any sufficiently small $\psi > 0$.

To see that the second claim in the proposition is true, consider some sample of size N , with A_s realizations of heads, and assume that $\theta_A > \frac{1}{2}$ and $A_s > N - A_s$. (The other cases have analogous proofs.) Tommy's likelihood ratio is

$$\frac{\frac{N!}{\theta_A!(N-\theta_A)!} \theta_A^{A_s} (1-\theta_A)^{N-A_s}}{\frac{N!}{\theta_A!(N-\theta_A)!} \theta_A^{N-A_s} (1-\theta_A)^{A_s}} = \left(\frac{\theta_A}{1-\theta_A}\right)^{2A_s-N},$$

which has $2A_s - N$ terms multiplied together. Barney's likelihood ratio is

$$\frac{\Gamma(\theta_A \psi + A_s)}{\Gamma((1-\theta_A)\psi + A_s)} \frac{\Gamma((1-\theta_A)\psi + N - A_s)}{\Gamma(\theta_A \psi + N - A_s)} = \frac{(\theta_A \psi + A_s - 1) \dots (\theta_A \psi + N - A_s)}{((1-\theta_A)\psi + A_s - 1) \dots ((1-\theta_A)\psi + N - A_s)}.$$

This also has $2A_s - N$ terms. Since by assumption $\frac{\theta_A \psi}{(1-\theta_A)\psi} = \frac{\theta_A}{1-\theta_A} > 1$, then for any $k > 0$,

$$\frac{\theta_A}{1-\theta_A} = \frac{\theta_A \psi}{(1-\theta_A)\psi} > \frac{\theta_A \psi + k}{(1-\theta_A)\psi + k} > 1,$$

which implies that

$$\left(\frac{\theta_A}{1-\theta_A}\right)^{2A_s-N} > \frac{(\theta_A \psi + A_s - 1) \dots (\theta_A \psi + N - A_s)}{((1-\theta_A)\psi + A_s - 1) \dots ((1-\theta_A)\psi + N - A_s)} > 1.$$

Hence Barney under-infers relative to Tommy.

Finally, consider two samples, one of size N with A_s a -signals, and the other of size $N' > N$ with A'_s a -signals. Suppose both samples have the same difference, d , between the number of a and b signals: $d = A_s - (N - A_s) = 2A_s - N$ and $d = 2A'_s - N'$. Since Tommy's likelihood ratio is $\left(\frac{\theta_A}{1-\theta_A}\right)^d$, it is the same for both samples. To consider Barney's inference, assume that $\theta_A > \frac{1}{2}$ and $d > 0$. (The other cases have analogous proofs.) Barney's likelihood ratio, stated above, has d terms for both samples. Since by assumption $\frac{\theta_A \psi}{(1-\theta_A)\psi} = \frac{\theta_A}{1-\theta_A} > 1$, then for any $k' > k > 0$,

$$\frac{\theta_A \psi + k}{(1-\theta_A)\psi + k} > \frac{\theta_A \psi + k'}{(1-\theta_A)\psi + k'} > 1.$$

Note that $A'_s > A_s$. It follows that each term in Barney's likelihood ratio for the sample of size N is larger than the corresponding term in Barney's likelihood ratio for the sample of size N' (e.g., for the first term, $\frac{\theta_A \psi + A_s - 1}{(1-\theta_A)\psi + A_s - 1} > \frac{\theta_A \psi + A'_s - 1}{(1-\theta_A)\psi + A'_s - 1}$). Hence while Barney infers in favor of state A in both cases (like Tommy does), Barney's change in beliefs is smaller from the larger sample. \square

Proposition 6. *Suppose Barney and Tommy have simple, piecewise-linear loss-averse preferences as specified in (4). Fix any gamble (θ, h, t) , paying off $h > 0$ with probability θ and $-t$ with probability $1 - \theta$, that is better than fair: $\theta h > (1 - \theta)t$. For any $\lambda \geq 1$, there is some $N' \geq 1$ such that if $N > N'$, then Tommy will accept N repetitions of the gamble. In contrast, for Barney there is some threshold level of loss aversion $\hat{\lambda} > 1$ such that: if $\lambda < \hat{\lambda}$, then there is some N' sufficiently large such that Barney will accept N repetitions of the gamble for all $N > N'$; and if $\lambda \geq \hat{\lambda}$, then there is some N'' sufficiently large such that Barney will reject N repetitions of the gamble for all $N > N''$.*

Proof: WLOG let the reference point be w_0 , and fix $\lambda \geq 1$. Denoting $G(z)$ as the distribution of monetary outcomes z , the expected utility of any lottery is:

$$\begin{aligned} & \int_0^\infty (w_0 + z) dG(z) + \int_{-\infty}^0 (w_0 + \lambda z) dG(z) \\ &= w_0 + \int_0^\infty z dG(z) + \lambda \int_{-\infty}^0 z dG(z). \end{aligned}$$

Clearly, this is better than the option of refusing the lottery if and only if

$$\int_0^\infty z dG(z) > -\lambda \int_{-\infty}^0 z dG(z) \Leftrightarrow \frac{\int_0^\infty z dG(z)}{-\int_{-\infty}^0 z dG(z)} < \lambda.$$

The monetary outcome of an N -times repeated gamble is $z = A_s h - (N - A_s)t = -Nt + A_s(h + t) = N(-t + \frac{A_s}{N}(h + t))$. Hence the agent earns a positive payoff from the gamble if and only if $A_s \geq \lceil \frac{t}{h+t}N \rceil$. Substituting in, we find that for Tommy,

$$\frac{\int_0^\infty z dG(z)}{-\int_{-\infty}^0 z dG(z)} = \frac{\sum_{A_s=\lceil \frac{t}{h+t}N \rceil}^N (-t + \frac{A_s}{N}(h+t)) f_{S_N|\Theta}(A_s|\theta)}{\lceil \frac{t}{h+t}N \rceil - 1 - \sum_{A_s=0}^{\lceil \frac{t}{h+t}N \rceil - 1} (-t + \frac{A_s}{N}(h+t)) f_{S_N|\Theta}(A_s|\theta)},$$

where θ is the known rate of the good outcome (and the N 's in the numerator and denominator cancel out).

Since the bet is better-than-fair, the probability of losing money goes to 0 as $N \rightarrow \infty$. Therefore, as $N \rightarrow \infty$, the denominator goes to 0. As $N \rightarrow \infty$, $\frac{A_s}{N}$ converges almost surely to θ , so the numerator goes to $-t + \theta(h + t) \geq 0$. Hence the ratio of the numerator to the denominator goes to infinity. This implies that Tommy always accepts the N -times repeated gamble for N sufficiently large.

Alternatively, for Barney,

$$\begin{aligned} \frac{\int_0^\infty z dG(z)}{-\int_{-\infty}^0 z dG(z)} &= \frac{\sum_{A_s=\lceil \frac{t}{h+t}N \rceil}^N (-t + \frac{A_s}{N}(h+t)) f_{S_N|\Theta}^\psi(A_s|\theta)}{\lceil \frac{t}{h+t}N \rceil - 1 - \sum_{A_s=0}^{\lceil \frac{t}{h+t}N \rceil - 1} (-t + \frac{A_s}{N}(h+t)) f_{S_N|\Theta}^\psi(A_s|\theta)} \\ &\rightarrow \frac{\int_{\frac{t}{h+t}}^1 (-t + \beta(h+t)) f_{B|\Theta}^\psi(\beta|\theta)}{-\int_0^{\frac{t}{h+t}} (-t + \beta(h+t)) f_{B|\Theta}^\psi(\beta|\theta)} \end{aligned}$$

almost surely as $N \rightarrow \infty$. This is a finite, positive number. □

Proposition 7. *Fix payoffs $u(\mu, \omega)$, rates $\theta_A > \theta_B$, and priors $f_\Theta(\theta_A)$ and $f_\Theta(\theta_B)$. Suppose Barney is prospective-atomizing or prospective-acceptive.*

1. *If Barney is retrospective-atomizing or retrospective-acceptive, then for any $c > 0$, Barney's behavior exactly coincides with Tommy's, and he will choose an action after observing a finite number of signals almost surely.*
2. *If Barney is retrospective-pooling, then for all $\bar{p} < 1$, there exists $\bar{c} > 0$ such that for all $c \leq \bar{c}$, Barney purchases an infinite number of signals with probability $p > \bar{p}$.*

3. Suppose Barney is retrospective-pooling, $\theta_A = 1 - \theta_B$, and Barney is willing to purchase an additional signal when his posterior probability of state A equal to q . (a) If Barney's posterior probability of state A is q after observing N signals and again q after observing $N' > N$ signals, then the probability that Barney will purchase an infinite number of signals is weakly higher after he has observed the N' signals than it was after the N signals. (b) Furthermore, there exists a \hat{N} such that for all $N' > \hat{N}$, the probability that Barney will purchase an infinite number of signals is strictly higher after he has observed the N' signals than after the N signals.

Proof: Since Barney is prospective-atomizing, he conceives of his problem as in the classic sequential information acquisition setting of Wald (1947). Therefore, we can characterize his optimal policy using two thresholds in the posterior ratio. Denote the lower and upper threshold by $\underline{\lambda}$ and $\bar{\lambda}$, respectively. These thresholds are functions of the cost of signals and the payoffs for each action in each state (but not on the current posterior ratio). So long as Barney's current posterior ratio is in $(\underline{\lambda}, \bar{\lambda})$, he will continue to acquire information. Once the posterior ratio is in the region $(0, \underline{\lambda}] \cup [\bar{\lambda}, 1)$, Barney will stop and take an action.

If Barney is retrospective-atomizing, then he is simply Tommy, who as it is well known, eventually stops in finite time almost surely (Wald, 1947). If Barney is retrospective-acceptive, then because he receives the signals one at a time, he will group them singly—and so his beliefs after any set of signals will be exactly the same as retrospective-atomizing Barney, and so the same results apply.

To prove the second part, now assume Barney is retrospective-pooling, and WLOG that state A is the true state. If Barney is retrospective-pooling, then for any two values m and n , such that $0 < m < n < \infty$, we can find a c' small enough so that if $c \leq c'$, then $\underline{\lambda} < m < n < \bar{\lambda}$. This is because, as in Wald (1947), the upper (lower) threshold is strictly increasing (decreasing) in c without bound.

Since $0 < p_0 < \infty$ and Barney's limit likelihood ratio $\Pi_{S_\infty|\Theta \times \Theta}^\psi (s_\infty|\theta_A, \theta_B, \Omega = i)$ is finite, his limit posterior is bounded away from 0 and ∞ :

$$0 < p_0 \Pi_{S_\infty|\Theta \times \Theta}^\psi (s_\infty|\theta_A, \theta_B, \Omega = i) < \infty.$$

Moreover, since the posterior ratios after just a single a -signal or a single b -signal are bounded away from 0 and ∞ , we can find a small enough $\hat{c} > 0$ so that both the following statements are true for all $c < \hat{c}$:

1. $[p_0 \Pi_{S_1|\Theta \times \Theta}^\psi (b|\theta_A, \theta_B), p_0 \Pi_{S_1|\Theta \times \Theta}^\psi (a|\theta_A, \theta_B)] \subseteq (\underline{\lambda}, \bar{\lambda})$

2. $p_0 \Pi_{S_\infty | \Theta \times \Theta}^\psi (s_\infty | \theta_A, \theta_B, \Omega = i) p_0 \in (\underline{\lambda}, \bar{\lambda})$.

Therefore, by the proof of Lemma C, if $c < c'' \equiv \min\{c', \hat{c}\}$, then the probability that retrospective-pooling Barney purchases an infinite number of signals, conditional on still purchasing signals after N signals have already been purchased, is increasing in N and converges to 1 as N goes to infinity.

Fix $\bar{p} < 1$. By the above argument, for any $c < c''$, we find an \bar{N} such that the probability that Barney purchases an infinite number of signals, conditional on Barney still purchasing signals after \bar{N} signals have already been purchased, is larger than \bar{p} . Now, there exist $0 < m < n < \infty$ sufficiently far apart such that the probability equals 1 that retrospective-pooling Barney's posterior ratio lies the region (m, n) after he observes \bar{N} signals. We can find c''' sufficiently small that for all $c \leq c'''$, $\underline{\lambda} < m < n < \bar{\lambda}$. It follows that if $c < \bar{c} \equiv \min\{c'', c'''\}$, then Barney purchases an infinite number of signals with probability larger than \bar{p} .

Turning to (a) of the third part of the proposition, denote the situations where there is a current posterior ratio of q , which is a belief where the agent still strictly wants to experiment, after N and $N' > N$ signals as q_N and $q_{N'}$, respectively. Assume that given $q_{N'}$, a sequence of m signals, denoted s_m , causes the agent to stop purchasing signals and (WLOG) take action a , and that no truncation of s_m causes the agent to stop purchasing signals. We will show that given initial situation q_N , the sequence s_m also causes the agent to stop purchasing signals. This proves that the probability of the agent stopping is weakly higher given q_N than $q_{N'}$.

If any truncation of s_m causes the agent to stop given q_N , then we are done, so assume not.

Now we shall show that a sufficient condition for part (a) is that the posterior induced by k a -signals in a row is less in favor of state A when starting with $q_{N'}$ than when starting with q_N . To see this, reorder s_m so that it begins with a and b signals paired off in alternating orders. Note that this reordering does not change the posterior after s_m , only the path of posterior ratios. Since by assumption the agent ends up favoring action a , there must be more a -signals than b -signals, so the reordered s_m must end with some set of a -signals in a row. Denote this number as k . Note that after every pair of signals, a, b , the likelihood ratio must still be q , since the signal structure is symmetric. So therefore, starting with either q_N or $q_{N'}$, after $m - k$ signals (the pairs of a and b signals) of the reordered s_m , the likelihood ratio is still q . Hence we are now simply comparing the effect of k a -signals in a row, given $N + m - k$ preceding signals versus $N' + m - k$ preceding signals.

Note that given $q_{N'}$, there must have been more preceding a -signals than given q_N . Assume there were r' and r a -signals, respectively. Hence in the sequence $q_{N'} \cup s_m$, there is a total of $N' + m$ signals, and the number of a -signals is r' (the number of a -signals in the first N') plus $\frac{m-k}{2}$ (the number of a -signals in the alternating set) plus k (the number of a -signals in a row at the end).

Therefore, using Lemma $\beta 2$, the agent's posterior ratio after $q_{N'} \cup s_m$ is

$$\begin{aligned}
& \Pi_{\Theta \times \Theta | S_{N'+m}}^\psi (\theta_A, \theta_B | q_{N'} \cup s_m) \\
&= p_0 \frac{\Gamma(\theta_A \psi + r' + \frac{m-k}{2} + k) \Gamma((1-\theta_A)\psi + N' - r' + \frac{m-k}{2})}{\Gamma(\theta_B \psi + r' + \frac{m-k}{2} + k) \Gamma((1-\theta_B)\psi + N' - r' + \frac{m-k}{2})} \\
&= p_0 \frac{\Gamma(\theta_A \psi + r' + \frac{m-k}{2}) \Gamma((1-\theta_A)\psi + N' - r' + \frac{m-k}{2})}{\Gamma(\theta_B \psi + r' + \frac{m-k}{2}) \Gamma((1-\theta_B)\psi + N' - r' + \frac{m-k}{2})} \prod_{i=1}^k \frac{\theta_A \psi + r' + \frac{m-k}{2} + i}{(1-\theta_A)\psi + r' + \frac{m-k}{2} + i} \\
&= q \prod_{i=1}^k \frac{\theta_A \psi + r' + \frac{m-k}{2} + i}{(1-\theta_A)\psi + r' + \frac{m-k}{2} + i}.
\end{aligned}$$

In the sequence $q_N \cup s_m$, there is a total of $N + m$ signals, and the number of a -signals is r (the number of a -signals in the first N) plus $\frac{m-k}{2}$ (the number of a -signals in the alternating set) plus k (the number of a -signals in a row at the end). Therefore, analogously, the agent's posterior ratio after $q_N \cup s_m$ is

$$\Pi_{\Theta \times \Theta | S_{N'+m}}^\psi (\theta_A, \theta_B | q_N \cup s_m) = q \prod_{i=1}^k \frac{\theta_A \psi + r + \frac{m-k}{2} + i}{(1-\theta_A)\psi + r + \frac{m-k}{2} + i}.$$

Notice that for all i ,

$$1 < \frac{\theta_A \psi + r' + \frac{m-k}{2} + i}{(1-\theta_A)\psi + r' + \frac{m-k}{2} + i} < \frac{\theta_A \psi + r + \frac{m-k}{2} + i}{(1-\theta_A)\psi + r + \frac{m-k}{2} + i}$$

since $r' > r$ and both fractions are larger than 1. Therefore, since the agent stops purchasing signals given his posterior in situation $q_{N'}$, he must stop given his posterior in situation q_N , which favors state A even more strongly.

To show part (b), note that as N and N' grow farther apart, so do r and r' . Fix N , and WLOG, assume that the agent stops purchasing signals after a sequence of all k a -signals and not before. Note that as $N' \rightarrow \infty$ with k fixed, $r' \rightarrow \infty$, and so the agent's likelihood ratio becomes increasingly dominated by the alternating sequence of a and b signals that precedes the k a -signals:

$$\prod_{i=1}^k \frac{\theta_A \psi + r' + \frac{m-k}{2} + i}{(1-\theta_A)\psi + r' + \frac{m-k}{2} + i} \rightarrow 1.$$

Therefore, the agent's posterior ratio approaches q ,

$$\prod_{i=1}^k \frac{\theta_A \psi + r' + \frac{m-k}{2} + i}{(1-\theta_A)\psi + r' + \frac{m-k}{2} + i} \rightarrow q,$$

and so the agent will *strictly* want to experiment for a large enough N' .

Proposition 8. *Suppose Barney is strategically sophisticated. Fix rates $\theta_A > \theta_B$ and Barneyness parameter $0 < \psi < \infty$. There are bounds $0 < \underline{p} < \frac{1}{2}$ and $\frac{1}{2} < \bar{p} < 1$ such that if the common prior $f_\Theta(\theta_A) \geq \underline{p}$, then there is positive probability of herding on action a ; and if $f_\Theta(\theta_A) \leq \bar{p}$, then there is positive probability of herding on action b .*

Proof: All types of agents, after observing the set of actions taken previously to them, and their own signal, will simply take the action optimal for them given their beliefs. Denoting the private posterior ratio of state A to state B by agent t , conditional on observing the vector of previous actions α_{t-1} and the private signal s as $\Pi_{\Theta \times \Theta | S_t}^\psi(\theta_A, \theta_B | \alpha_{t-1}, s)$, an agent will take action a if $\Pi_{\Theta \times \Theta | S_t}^\psi(\theta_A, \theta_B | \alpha_{t-1}, s) > 1$ and action b if $\Pi_{\Theta \times \Theta | S_t}^\psi(\theta_A, \theta_B | \alpha_{t-1}, s) < 1$. As noted in the text, we focus on the most informative equilibrium: if $\Pi_{\Theta \times \Theta | S_t}^\psi(\theta_A, \theta_B | \alpha_{t-1}, s) = 1$, then an agent will take the action that matches his own signal s .

Denote the common prior ratio as p_0 . We will show that if $p_0 \geq \underline{p} \equiv \frac{1}{\Pi_{S_1 | \Theta \times \Theta}^\psi(a | \theta_A, \theta_B)}$, then there is positive probability that parameterized-Barney will herd on action a , and if $p_0 \leq \bar{p} \equiv \frac{1}{\Pi_{S_1 | \Theta \times \Theta}^\psi(b | \theta_A, \theta_B)}$, then there is positive probability that parameterized-Barney will herd on action b . Proving this claim will prove the proposition.

First, consider a common prior ratio in the interval:

$$p_0 \in [0, \infty) \setminus \left[\frac{1}{\Pi_{S_1 | \Theta \times \Theta}^\psi(a | \theta_A, \theta_B)}, \frac{1}{\Pi_{S_1 | \Theta \times \Theta}^\psi(b | \theta_A, \theta_B)} \right].$$

If $p_0 \leq \frac{1}{\Pi_{S_1 | \Theta \times \Theta}^\psi(a | \theta_A, \theta_B)}$, it will be optimal for the first agent to take action b regardless of his signal, and so a herd on action b begins in period 1. If $p_0 \geq \frac{1}{\Pi_{S_1 | \Theta \times \Theta}^\psi(b | \theta_A, \theta_B)}$, it will be optimal for the first agent to take action a regardless of his signal, and so a herd on action a begins in period 1.

Now consider the set of prior ratios for which no herd begins in period 1:

$$p_0 \in \left[\frac{1}{\Pi_{S_1 | \Theta \times \Theta}^\psi(a | \theta_A, \theta_B)}, \frac{1}{\Pi_{S_1 | \Theta \times \Theta}^\psi(b | \theta_A, \theta_B)} \right].$$

We will conclude the proof by showing that, for prior ratios in this interval, there is positive probability of a herd occurring on either action. Note that for any number $L \in (1, \infty)$, there exists a long enough sequence of all a -signals such that parameterized-Barney's likelihood ratio is $\geq L$ after observing this sequence (due to the functional form of the beta distribution). Given that p_0 is in the above interval, it follows that for a large enough sample of signals, there is positive probability that beta-binomial-Barney's posterior ratio will be larger than $\frac{1}{\Pi_{S_1 | \Theta \times \Theta}^\psi(b | \theta_A, \theta_B)}$, and so a herd will begin on action a . An analogous argument shows that there is positive probability that

a herd will begin on action b .

□

Proposition 9. *Suppose Barney is strategically sophisticated. There exist $\delta > 0$ and $\gamma > 0$ such that if $|f_{\Theta}(\theta_A) - f_{\Theta}(\theta_B)| < \delta$ and $|\theta_A - \theta_B| < \gamma$, then there is positive probability of an eddy regardless of which state is true.*

Proof: Strategically sophisticated Barney believes that a herd has started if and only if a herd has actually started. If a herd has started, Barney will ignore his own signal and join the herd, since the previous agent faced the same problem and joined the herd, and Barney has accurate beliefs about that agent's beliefs. So we only need to consider situations where Barney believes no herd has started.

For an eddy to exist, we need that agents will never ignore their own signal. As long as the public posterior ratio is in the interval

$$\left[\frac{1}{\Pi_{S_1|\Theta \times \Theta}^{\psi}(a|\theta_A, \theta_B)}, \frac{1}{\Pi_{S_1|\Theta \times \Theta}^{\psi}(b|\theta_A, \theta_B)} \right],$$

agents will take actions corresponding to their signal. This is because the agent's private posterior ratio after observing his own signal will be between

$$\frac{\Pi_{S_1|\Theta \times \Theta}^{\psi}(b|\theta_A, \theta_B)}{\Pi_{S_1|\Theta \times \Theta}^{\psi}(a|\theta_A, \theta_B)}$$

and

$$\frac{\Pi_{S_1|\Theta \times \Theta}^{\psi}(a|\theta_A, \theta_B)}{\Pi_{S_1|\Theta \times \Theta}^{\psi}(b|\theta_A, \theta_B)},$$

an interval that includes 1.

WLOG, assume that the true state is A . To guarantee that there is positive probability that the public posterior ratio remains in the above interval, it is sufficient that the prior and limit likelihood ratio satisfy two conditions:

- The first agent must not ignore his own signal; this means that

$$p_0 \in \left[\frac{1}{\Pi_{S_1|\Theta \times \Theta}^{\psi}(a|\theta_A, \theta_B)}, \frac{1}{\Pi_{S_1|\Theta \times \Theta}^{\psi}(b|\theta_A, \theta_B)} \right].$$

- The limit posterior conditional on the true state is:

$$\Pi_{S_\infty|\Theta \times \Theta}^\psi (s_\infty|\theta_A, \theta_B, \Omega = A) p_0 \in \left[\frac{1}{\Pi_{S_1|\Theta \times \Theta}^\psi (a|\theta_A, \theta_B)}, \frac{1}{\Pi_{S_1|\Theta \times \Theta}^\psi (b|\theta_A, \theta_B)} \right].$$

Since

$$1 \in \left[\frac{1}{\Pi_{S_1|\Theta \times \Theta}^\psi (a|\theta_A, \theta_B)}, \frac{1}{\Pi_{S_1|\Theta \times \Theta}^\psi (b|\theta_A, \theta_B)} \right],$$

we can find some δ such that if $|f_\Theta(\theta_A) - f_\Theta(\theta_B)| < \delta$, then the first condition holds. Furthermore, since the limit likelihood ratio, $\Pi_{S_\infty|\Theta \times \Theta}^\psi (s_\infty|\theta_A, \theta_B, \Omega = A)$, is equal to 1 when $\theta_A = \theta_B$, and is continuous in both rates, it follows that for any $\tau > 0$ there exists a γ such that if $|\theta_A - \theta_B| < \gamma$, then the limit likelihood ratio (conditional on the true state) is within τ of 1. Therefore, conditional on $|f_\Theta(\theta_A) - f_\Theta(\theta_B)| < \delta$, we can find rates θ_A and θ_B such that the second condition is also satisfied.

We can now use Lemma C. Although Lemma C as stated applies to retrospective-pooling Barney, we can use it here to bound the public posterior likelihood ratio. This is because Barney is retrospective-pooling with respect to the public signals he has inferred from previous actions. Therefore, we know that with positive probability, the public posterior ratio stays within $\left[\frac{1}{\Pi_{S_1|\Theta \times \Theta}^\psi (a|\theta_A, \theta_B)}, \frac{1}{\Pi_{S_1|\Theta \times \Theta}^\psi (b|\theta_A, \theta_B)} \right]$ forever. Therefore, with positive probability, an eddy will result. \square

Proposition 10. *Suppose Barney is strategically sophisticated.*

1. For any $\varepsilon > 0$, there exists $T_\varepsilon > 2$ such that if a herd has not occurred by period T_ε , the probability of herding from then on is less than ε .
2. Furthermore, suppose the true state is A , the prior $f_\Theta(\theta_A) \geq .5$, and $1 - \theta_A = \theta_B$. For any $\nu > 0$ there exists $T_\nu > 2$ such that if a herd has not occurred by period T_ν , the likelihood ratio of a bad herd to a good herd is less than ν .

Proof:

1. For a herd to begin, the public posterior ratio must leave $\left[\Pi_{S_1|\Theta \times \Theta}^\psi (b|\theta_A, \theta_B), \Pi_{S_1|\Theta \times \Theta}^\psi (a|\theta_A, \theta_B) \right]$. Since an eddy occurs with positive probability, from the proof of Lemma C, we know that if a herd has not yet occurred after \hat{N} signals, the probability of the public posterior ratio leaving $\left[\Pi_{S_1|\Theta \times \Theta}^\psi (b|\theta_A, \theta_B), \Pi_{S_1|\Theta \times \Theta}^\psi (a|\theta_A, \theta_B) \right]$ is less than $1 - \sum_{i=\hat{N}}^\infty \frac{p(1-p)}{i^2 \delta^2}$, where $p = \Pi_{S_1|\Theta \times \Theta}^\psi (b|\theta_A, \theta_B)$ and δ is the distance between the limit proportion of a -signals and the

closest proportion that induces a herd. This expression is an upper bound on the probability of a herd ever starting, if it has not already started after \hat{N} signals have occurred. As noted in the proof of Lemma C, this upper bound converges to 0 as $\hat{N} \rightarrow \infty$; hence, for any $\epsilon > 0$, we can find a T_ϵ such that if a herd has not occurred by period T_ϵ , then the probability of a herd occurring is less than ϵ .

2. Denote the prior ratio on state A by $p = \frac{f_\Theta(\theta_A)}{f_\Theta(\theta_B)}$. Denote the probability of a good herd (that is, a herd on state A) given parameters $\theta_A, \theta_B, \psi, p$, and with N agents already have taken an action, as $V(A, \theta_A, \theta_B, \psi, N, p)$. Denote the probability of a bad herd (a herd on state B) as $V(B, \theta_A, \theta_B, \psi, N, p)$. For brevity, we will refer to these as $V(A, N, p)$ and $V(B, N, p)$, respectively. We will show that, for any $\nu > 0$, there exists $T_\nu > 2$ such that if a herd has not occurred by period T_ν , then for all $N \geq T_\nu$, $\frac{V(B, N, p)}{V(A, N, p)} < \nu$. We will do this in several steps:

- (a) First, we will construct $V(i, N, p)$ in a way that will make it manageable to work with.
- (b) Second, we will fix equal priors $p = 1$, and we will show that, for any $\nu > 0$, there exists $T_\nu > 2$ such that for all $N \geq T_\nu$, $\frac{V(B, N, 1)}{V(A, N, 1)} < \nu$.
- (c) Third, we will show that for any priors $p > 1$ that strictly favor the true state and for all $N \geq T_\nu$, $\frac{V(B, N, p)}{V(A, N, p)} < \frac{V(B, N, 1)}{V(A, N, 1)}$. This is because relative to the equal priors case, there is a uniform shift of probability from likelihoods that favor an incorrect herd to a correct herd.

Now we shall do each step in order:

- (a) Because the signals are symmetric, we can define

$$\lambda \equiv \frac{1}{\Pi_{S_1|\Theta \times \Theta}^\psi(a|\theta_A, \theta_B)} = \Pi_{S_1|\Theta \times \Theta}^\psi(b|\theta_A, \theta_B).$$

The set of public likelihood ratios such that a herd has not yet started after N signals is denoted $Q(N) \subset [\lambda, \frac{1}{\lambda}]$, a set with a finite number of elements. Denote a typical element by q .

We will construct V from a more elementary function: $\phi(A, q, N, M)$, the probability of a herd starting on the a action in the next M signals when the current public likelihood ratio is q , and N signals have already been observed. Given p and N , there is a one-to-one mapping between the public likelihood ratio q and the “number of a signals” previously observed, which may not be a natural number for a particular q . Therefore we will consider the number of a and b signals as fixed at the levels implied by q .

After N agents have taken actions, a herd starts on action a only if the number of a -signals exceeds the number of b -signals by at least some number. Abusing notation slightly, denote by s_M^j an exact sequence of a and b signals from a set of M signals, with ordering j indexed from 0 to $2^M - 1$ (so s_M^j could contain any number of signals from 0 up to M signals). Denote the number of a -signals in s_M^j as $A(s_M^j)$. Define $a(s_M^j, N, q) \in \{0, 1\}$ as an indicator of whether a herd would have started on action a with the exact ordering of signals in s_M^j , given that the public likelihood ratio was q after the first N signals. Now we can construct the probability of a herd starting on action a within the next M signals, given that the public likelihood ratio was q after the first N signals:

$$\phi(A, q, N, M) = \sum_{j=1}^{2^M-1} a(s_M^j, N, q) \theta_A^{A(s_M^j)} (1 - \theta_A)^{M-A(s_M^j)},$$

a polynomial in θ_A and $1 - \theta_A$.

To gain some intuition for ϕ , consider $q = 1$. Because the rates are symmetric, if we consider a set of j signals that causes a herd on the a action, and we replace all the θ_A 's with $(1 - \theta_A)$'s, and vice-versa, then that set of signals will cause a herd on action b . Similarly, if we have a set of j signals that has not caused a herd, then the same set of signals, but replacing θ_A 's with $(1 - \theta_A)$'s and vice-versa, will also not cause a herd. Therefore, $\phi(B, 1, N, M)$ is the same function as $\phi(A, 1, N, M)$ but with all the θ_A 's replaced with $(1 - \theta_A)$'s and vice-versa. Now suppose either $q > 1$ or $q < 1$. Now $\phi(B, q, N, M)$ is not simply $\phi(A, q, N, M)$ with the rates switched. Because rates are symmetric, however, if we define $q' = \frac{1}{q}$, then $\phi(B, q', N, M)$ is the same function as $\phi(A, q, N, M)$ but with θ_A and $1 - \theta_A$ switched.

Define $\rho(N)$ as the minimum number of a -signals in a row that would need to occur after an initial set of N signals to put the public likelihood ratio above $\frac{1}{\lambda}$ (the upper bound of the interval $Q(N)$, as defined above), given that the public likelihood ratio was λ (the lower bound of the interval $Q(N)$) after the first N signals. $\rho(N)$ will be a useful function to define because, regardless of the public likelihood ratio after N observations, it is the number of additional signals after which a herd on either action could always possibly start. Note that $\rho(N)$ is increasing in N .

Define $x(q, N, p) < N$ as the number of a -signals out of N such that $x(q, N, p)$ a -signals and $N - x(q, N, p)$ b -signals induces posterior ratio of q when the prior is p .

For each $x(q, N, p)$, we can construct the set of possible sequences that generates x a -signals and $N - x$ b -signals without starting a herd at any point in the sequence. Denote the set of these sequences as $X(q, N, p)$. Each one of these sequences must have the same final number of a and b signals, but it must be the case that the difference between the number of a -signals and b -signals cannot have been too large at any point in the sequence, or else a herd would have started. Let

$g(q, N, p) = \gamma(q, N, p)\theta_A^{x(q, N, p)}(1 - \theta_A)^{N-x(q, N, p)}$ denote the probability of the set of sequences that leads to x a -signals and $N - x$ b -signals without a herd ever having started, where γ is the number of sequences in the set.

Define $r_N(i)$ for $i = 0, 1, \dots$ recursively: $r_N(0) = N$ and $r_N(i) = r_N(i - 1) + \rho(r_N(i - 1))$ for $i > 0$.

We now construct $V(A, N, p)$ as follows:

$$V(A, N, p) = \sum_{i=1}^{\infty} \sum_{q \in Q(r_N(i-1))} \phi(A, q, r_N(i-1), \rho(r_N(i-1)))g(q, r_N(i-1), p).$$

We define $V(B, N, p)$ analogously.

To understand this definition of V , each term is the probability of a herd starting on the appropriate action within the next $r_N(i)$ signals, conditional on having observed $r_N(i - 1)$ signals without a herd, starting with a prior ratio of p .

(b) Now we will fix the prior ratio at 1 and show that for any $\nu > 0$, there exists $T_\nu > 2$ such that for all $N \geq T_\nu$, $\frac{V(B, N, 1)}{V(A, N, 1)} < \nu$.

We will begin by proving some key claims:

Claim 10.A Let $a_i, b_i > 0$ for $i = 1, 2, \dots$, such that $\frac{a_i}{b_i} \leq \frac{a_{i+1}}{b_{i+1}}$. Then $\frac{a_1}{b_1} \leq \frac{\sum_{i=1}^N a_i}{\sum_{i=1}^N b_i} \leq \frac{a_N}{b_N}$.

To see this is true, note that for positive a, b, c, d , if $0 < \frac{a}{b} < \frac{c}{d}$, then $\frac{a}{b} < \frac{a+c}{b+d} < \frac{c}{d}$. Induction then shows that $\frac{\sum_{i=1}^N a_i}{\sum_{i=1}^N b_i}$ is bounded by the maximum and minimum $\frac{a_i}{b_i}$.

Claim 10.B For any $\nu > 0$, there exists $T_\nu > 2$ such that for all $N \geq T_\nu$,

$$\frac{\sum_{q \in Q(N)} \phi(B, q, N, \rho(N))g(q, N, p)}{\sum_{q \in Q(N)} \phi(A, q, N, \rho(N))g(q, N, p)} < \nu.$$

Together, Claims 10.A and 10.B imply the result we are seeking. Claim 10.A, combined with the fact that the probability of a herd is shrinking as the number of signals gets large, implies that

$$\begin{aligned} \frac{V(B, N, p)}{V(A, N, p)} &= \frac{\sum_{i=1}^{\infty} \sum_{q \in Q(N)} \phi(B, q, r_N(i-1), r_N(i))g(q, r_N(i-1), p)}{\sum_{i=1}^{\infty} \sum_{q \in Q(N)} \phi(A, q, r_N(i-1), r_N(i))g(q, r_N(i-1), p)} \\ &\leq \frac{\sum_{q \in Q(N)} \phi(B, q, N, \rho(N))g(q, N, p)}{\sum_{q \in Q(N)} \phi(A, q, N, \rho(N))g(q, N, p)}. \end{aligned}$$

Claim 10.B then says that given any $\nu > 0$, for large enough N , this ratio is $< \nu$.

A sufficient condition for Claim 10.B is:

Claim 10.C For any $\nu > 0$, there exists $T_\nu > 2$ such that for all $N \geq T_\nu$ and $q \in Q(N)$,

$$\frac{\phi(B, q, N, \rho(N))g(q, N, 1) + \phi(B, q', N, \rho(N))g(q', N, 1)}{\phi(A, q, N, \rho(N))g(q, N, 1) + \phi(A, q', N, \rho(N))g(q', N, 1)} < \nu.$$

To see that Claim 10.C implies Claim 10.B, notice that $\frac{\sum_{q \in Q(N)} \phi(B, q, N, \rho(N))g(q, N, p)}{\sum_{q \in Q(N)} \phi(A, q, N, \rho(N))g(q, N, p)}$ is

$$\begin{aligned} &\leq \max_{q, q' \in Q(N) \text{ s.t. } q' = \frac{1}{q}} \frac{\phi(B, q, N, \rho(N))g(q, N, 1) + \phi(B, q', N, \rho(N))g(q', N, 1)}{\phi(A, q, N, \rho(N))g(q, N, 1) + \phi(A, q', N, \rho(N))g(q', N, 1)} \\ &< \nu \end{aligned}$$

where the first inequality follows from Claim 10.A and the second by Claim 10.C.

Therefore, we will have completed part (b) once we prove Claim 10.C. We substitute:

$$\begin{aligned} &\phi(A, q, N, \rho(N))g(q, N, 1) \\ &= g(q, N, 1) \sum_{j=1}^{2^{\rho(N)}-1} a(s_{\rho(N)}^j, N, q) \theta_A^{A(s_{\rho(N)}^j)} (1 - \theta_A)^{\rho(N) - A(s_{\rho(N)}^j)} \\ &= \gamma(q, N, 1) \theta_A^{x(q, N, 1)} (1 - \theta_A)^{N - x(q, N, 1)} \sum_{j=1}^{2^{\rho(N)}-1} a(s_{\rho(N)}^j, N, q) \theta_A^{A(s_{\rho(N)}^j)} (1 - \theta_A)^{\rho(N) - A(s_{\rho(N)}^j)}. \end{aligned}$$

Note that some of the sequences of signals starting with a posterior of q will cause a herd before $\rho(N)$ signals. Imagine two sequences begin with the same initial z signals but differ in the signals that occur afterward. If the z^{th} signal begins a herd, then these subsequent signals are irrelevant, and we can aggregate the two sequences together. Let $\zeta(q, \rho(N), N, z, i)$ denote the number of a -signals needed out of an additional $\rho(N)$ signals, starting with a posterior of q after N signals, for a herd on i to begin after exactly z signals, and no herd having started sooner. Let the lowercase letter ς count the number of permutations of such sequences. Continuing to substitute:

$$\begin{aligned} &\phi(A, q, N, \rho(N))g(q, N, 1) \\ &= \gamma(q, N, 1) \theta_A^{x(q, N, 1)} (1 - \theta_A)^{N - x(q, N, 1)} \sum_{z=1}^{\rho(N)} \varsigma(q, \rho(N), N, z, A) \theta_A^{\zeta(q, z, N, A)} (1 - \theta_A)^{z - \zeta(q, z, N, A)}. \end{aligned}$$

Now, define

$$\begin{aligned} b_z(q', N) &\equiv \gamma(q', N, 1) \theta_A^{x(q', N, 1)} (1 - \theta_A)^{N - x(q', N, 1)} \varsigma(q', \rho(N), N, z, B) (1 - \theta_A)^{z - \zeta(q', z, N, B)} \theta_A^{\zeta(q', z, N, B)} \\ a_z(q', N) &\equiv \gamma(q, N, 1) \theta_A^{x(q, N, 1)} (1 - \theta_A)^{N - x(q, N, 1)} \varsigma(q, \rho(N), N, z, A) (1 - \theta_A)^{z - \zeta(q, z, N, A)} \theta_A^{\zeta(q, z, N, A)}. \end{aligned}$$

We can write

$$\frac{\phi(B, q, N, \rho(N))g(q, N, 1) + \phi(B, q', N, \rho(N))g(q', N, 1)}{\phi(A, q, N, \rho(N))g(q, N, 1) + \phi(A, q', N, \rho(N))g(q', N, 1)} = \frac{\sum_{z=1}^{\rho(N)} b_z(q, N) + \sum_{z=1}^{\rho(N)} b_z(q', N)}{\sum_{z=1}^{\rho(N)} a_z(q, N) + \sum_{z=1}^{\rho(N)} a_z(q', N)}$$

Starting from a prior ratio of 1, because the signals are symmetric, if x a -signals and $N - x$ b -signals leaves an agent with a posterior of q , then x b -signals and $N - x$ a -signals leaves an agent with a posterior of q' . Therefore $x(q', N, 1) = N - x(q, N, 1)$. Similarly, $\gamma(q', N, 1) = \gamma(q, N, 1)$. Also, starting with a posterior of q , if x additional a -signals and $N - x$ additional b -signals leaves an agent willing to herd on action A , then starting with a posterior of q' , x additional b -signals and $N - x$ additional a -signals leaves an agent willing to herd on action B . Therefore $\zeta(q', z, N, B) = z - \zeta(q, z, N, A)$ and $\zeta(q', \rho(N), N, z, B) = \zeta(q, \rho(N), N, z, A)$.

Using these relationships,

$$\begin{aligned} \frac{b_z(q', N)}{a_z(q', N)} &= \frac{\gamma(q', N, 1)\theta_A^{x(q', N, 1)}(1 - \theta_A)^{N-x(q', N, 1)}\zeta(q', \rho(N), N, z, B)(1 - \theta_A)^{z-\zeta(q', z, N, B)}\theta_A^{\zeta(q', z, N, B)}}{\gamma(q, N, 1)\theta_A^{x(q, N, 1)}(1 - \theta_A)^{N-x(q, N, 1)}\zeta(q, \rho(N), N, z, A)(1 - \theta_A)^{z-\zeta(q, z, N, A)}\theta_A^{\zeta(q, z, N, A)}} \\ &= \frac{\gamma(q, N, 1)\theta_A^{N-x(q, N, 1)}(1 - \theta_A)^{x(q, N, 1)}\zeta(q, \rho(N), N, z, A)(1 - \theta_A)^{\zeta(q, z, N, A)}\theta_A^{z-\zeta(q, z, N, A)}}{\gamma(q, N, 1)\theta_A^{x(q, N, 1)}(1 - \theta_A)^{N-x(q, N, 1)}\zeta(q, \rho(N), N, z, A)(1 - \theta_A)^{z-\zeta(q, z, N, A)}\theta_A^{\zeta(q, z, N, A)}}, \end{aligned}$$

which is well-defined as long as $\gamma(q, N, 1)\zeta(q, \rho(N), N, z, A) \neq 0$. (If $\gamma(q, N, 1)\zeta(q, \rho(N), N, z, A) = 0$, then b_z drops out from $V(B, N, p)$, and a_z drops out from $V(A, N, p)$. Hence these terms can be ignored.) Canceling and collecting terms:

$$\frac{b_z(q', N)}{a_z(q', N)} = \frac{(1 - \theta_A)^{x(q, N, 1) + \zeta(q, z, N, A)} (\theta_A)^{N+z-x(q, N, 1) - \zeta(q, z, N, A)}}{(\theta_A)^{x(q, N, 1) + \zeta(q, z, N, A)} (1 - \theta_A)^{N+z-x(q, N, 1) - \zeta(q, z, N, A)}}$$

Using Claim 10.A ,

$$\frac{\sum_z b_z(q', N)}{\sum_z a_z(q', N)} < \max_{z \text{ s.t. } \gamma(q, N, 1)\zeta(q', \rho(N), N, z, B) \neq 0} \frac{(1 - \theta_A)^{x(q, N, 1) + \zeta(q, z, N, A)} (\theta_A)^{N+z-x(q, N, 1) - \zeta(q, z, N, A)}}{(\theta_A)^{x(q, N, 1) + \zeta(q, z, N, A)} (1 - \theta_A)^{N+z-x(q, N, 1) - \zeta(q, z, N, A)}}.$$

The numerator is the probability that, starting with a prior ratio of 1, there will be a sequence of a and b signals, $N + z$ in total, that lands the agent just inside the zone where the agent will herd on action B . The denominator is the probability that, starting with a prior ratio of 1, there will be a sequence of a and b signals, $N + z$ in total, that lands the agent just inside the zone where the

agent will herd on action A . This expression can be rearranged to yield

$$\left(\frac{1 - \theta_A}{\theta_A}\right)^{2x(q, N, 1) + 2\zeta(q, z, N, A) - N - z},$$

where the exponent, $2x(q, N, 1) + 2\zeta(q, z, N, A) - N - z \equiv \delta(N + z)$, is the minimum difference between the number of a and b signals out of $N + z$ signals in total at which a herd will start; if there are $\delta(N + z)$ more a than b signals, then there will be a herd on A ; and if there are $\delta(N + z)$ more b than a signals, then there will be a herd on B . Notice that $\delta(N + z)$ is not a function of q . Because Barney's likelihood ratio asymptotically depends on the proportion of a signals, not the difference between the number of a and b signals, $\lim_{N+z \rightarrow \infty} \delta(N + z) = \infty$. Furthermore, since $\rho(N)$ is growing in N , then it follows that for any $\epsilon > 0$ we can find an N' large enough such that for all $q \in \left(\Pi_{S_1|\Theta \times \Theta}^\psi(b|\theta_A, \theta_B), \Pi_{S_1|\Theta \times \Theta}^\psi(a|\theta_A, \theta_B)\right)$ (and so for any $q \in Q$ for any N) and all $N > N'$,

$$\left(\frac{1 - \theta_A}{\theta_A}\right)^{2x(q, N, 1) + 2\zeta(q, z, N, A) - N - z},$$

is less than ϵ . It then follows directly from the previous claims that the upper bound on $\frac{V(B, N, p)}{V(A, N, p)}$, $\left(\frac{1 - \theta_A}{\theta_A}\right)^{\delta(N+z)}$, goes to 0 as $N \rightarrow \infty$.

Therefore, although Q varies with N , it must be the case that for any q in Q , the limit result holds. That is, the convergence is uniform over $q \in Q$.

(c) Now assume that the prior rate favors A : $p > 1$. Note that nothing in the construction of the ϕ function changes. However, because p is now closer to $q > 1$ than it is to $q' < 1$, $\frac{g(q, N, p)}{g(q', N, p)} < \frac{g(q, N, 1)}{g(q', N, 1)}$. Using this fact, and the result proven above for the $p = 1$ case: for any $\nu > 0$, there exists $T_\nu > 2$ such that for all $N \geq T_\nu$, for any $q \in Q(N)$,

$$\frac{\phi(B, q, N, \rho(N))g(q, N, p)}{\phi(A, q', N, \rho(N))g(q', N, p)} < \frac{\phi(B, q, N, \rho(N))g(q, N, 1)}{\phi(A, q', N, \rho(N))g(q', N, 1)} < \nu,$$

and

$$\frac{\phi(B, q', N, \rho(N))g(q', N, p)}{\phi(A, q, N, \rho(N))g(q, N, p)} < \frac{\phi(B, q', N, \rho(N))g(q', N, 1)}{\phi(A, q, N, \rho(N))g(q, N, 1)} < \nu.$$

We argued in part (a) that Claim 10.A implies

$$\frac{V(B, N, p)}{V(A, N, p)} \leq \frac{\sum_{q \in Q(N)} \phi(B, q, N, \rho(N))g(q, N, p)}{\sum_{q \in Q(N)} \phi(A, q, N, \rho(N))g(q, N, p)}.$$

Using Claim 10.A again, along with the above inequalities, gives the result:

$$\frac{V(B, N, p)}{V(A, N, p)} < \nu.$$

□

References

- Aghion, P., P. Bolton, C. Harris, and B. Julien**, “Optimal Learning by Experimentation,” *Review of Economic Studies*, 1991, 58, 621–654.
- Ali, S. Nageeb**, “Learning Self-Control,” *Quarterly Journal of Economics*, 2011, 126, 857–893.
- Banerjee, Abhijit V.**, “A Simple Model of Herd Behavior,” *Quarterly Journal of Economics*, 1992, 107 (3), 797–817.
- Bar-Hillel, Maya**, “The base-rate fallacy in probability judgments,” *Acta Psychologica*, 1980, 44 (3), 211–233.
- Barberis, Nicholas, Andrei Schleifer, and Robert Vishny**, “A model of investor sentiment,” *Journal of Financial Economics*, 1998, 49 (3), 307–343.
- , **Ming Huang, and Richard H. Thaler**, “Individual Preferences, Monetary Gambles, and Stock Market Participation: A Case for Narrow Framing,” *American Economic Review*, 2006, 96 (4), 1069–1090.
- Beach, Lee Roy**, “Probability Magnitudes and Conservative Revision of Subjective Probabilities,” *Journal of Experimental Psychology*, 1968, 77 (1), 57–63.
- , **James A. Wise, and Scott Barclay**, “Sample Proportions and Subjective Probability Revisions,” *Organizational Behavior and Human Performance*, March 1970, 5 (2), 183–190.
- Benartzi, Shlomo and Richard Thaler**, “Myopic Loss Aversion and the Equity Premium Puzzle,” *The Quarterly Journal of Economics*, 1995, 110 (1), 73–92.
- Bernatzi, Shlomo and Richard Thaler**, “Risk Aversion or Myopia? Choices in Repeated Gambles and Retirement Investments,” *Management Science*, 1999, 45 (3), 364–381.
- Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch**, “A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades,” *Journal of Political Economy*, 1992, 100 (5), 992–1026.

- Blackwell, D.**, “Comparison of Experiments,” in “Proceedings of the Second Berkeley Symposium on Mathematics, Statistics and Probability,” University of California Press, 1951, pp. 93–102.
- , “Equivalent Comparisons of Experiments,” *Annals of Mathematical Statistics*, 1953, *24*, 265–272.
- Bodoh-Creed, Aaron L.**, “The Economics of Base Rate Neglect,” June 2010. Working paper.
- Camerer, Colin F.**, “Do Biases in Probability Judgment Matter in Markets? Experimental Evidence,” *The American Economic Review*, December 1987, *77* (5), 981–997.
- Chapman, Richard C.**, “Prior Probability Bias in Information Seeking and Opinion Revision,” *American Journal of Psychology*, 1973, *86* (2), 269–282.
- Chinnis, James O. Jr. and Cameron R. Peterson**, “Inference About a Nonstationary Process,” *Journal of Experimental Psychology*, 1968, *77* (4), 620–625.
- Cohen, John and C. E. M. Hansel**, “The Idea of a Distribution,” *British Journal of Psychology*, 1955, *46* (2), 111–121.
- Dale, H. C. A.**, “Weighing Evidence: An Attempt to Assess the Efficiency of the Human Operator,” *Ergonomics*, May 1968, *11* (3), 215–230.
- Dave, Chetan and Katherine W. Wolfe**, “On Confirmation Bias and Deviations From Bayesian Updating,” March 2003. Working paper.
- DeSwart, J.H.**, “A Comment on Marks and Clarkson’s Explanation of Conservatism,” *Acta Psychologica*, November 1972, *36* (5), 417–419.
- Donnell, Michael L. and Wesley M. Du Charmé**, “The Effect of Bayesian Feedback on Learning in an Odds Estimation Task,” *Organizational Behavior and Human Performance*, December 1975, *14* (3), 305–313.
- DuCharme, Wesley M.**, “Response Bias Explanation of Conservative Human Inference,” *Journal of Experimental Psychology*, 1970, *85* (1), 66–74.
- **and Cameron R. Peterson**, “Intuitive Inference About Normally Distributed Populations,” *Journal of Experimental Psychology*, 1968, *78* (2), 269–275.
- Edwards, W.**, “Conservatism in human information processing,” in B. Kleinmuntz, ed., *Formal representation of human judgment*, Wiley, 1968.

- Eyster, Erik and Matthew Rabin**, “Naive Herding in Rich-Information Settings,” *American Economic Journal: Microeconomics*, 2010, *2* (4), 221–243.
- Fried, Lisbeth S. and Cameron R. Peterson**, “Information Seeking: Optimal Versus Fixed Stopping,” *Journal of Experimental Psychology*, 1969, *80* (3), 525–529.
- Gettys, Charles F. and Charles W. Manley**, “The probability of an event and estimates of posterior probability based upon its occurrence,” *Psychonomic Science*, 1968, *11* (2), 47–48.
- Green, Paul E., Michael H. Halbert, and Patrick J. Robinson**, “An Experiment in Probability Estimation,” *Journal of Marketing Research*, August 1965, *2* (3), 266–273.
- Grether, David M.**, “Bayes Rule as a Descriptive Model: The Representativeness Heuristic,” *Quarterly Journal of Economics*, 1980, *95* (3), 537–557.
- , “Testing Bayes rule and the representativeness heuristic: Some experimental evidence,” *Journal of Economic Behavior and Organization*, January 1992, *17* (1), 31–57.
- Griffin, Dale and Amos Tversky**, “The Weighing of Evidence and the Determinants of Confidence,” *Cognitive Psychology*, July 1992, *24* (3), 411–435.
- Gustafson, David H., Ramesh K. Shukla, Andre Delbecq, and G. William Walster**, “A Comparative Study of Differences in Subjective Likelihood Estimates Made by Individuals, Interacting Groups, Delphi Groups, and Nominal Groups,” *Organizational Behavior And Human Performance*, April 1973, *9* (2), 280–291.
- Hershman, Ramon L. and J.R. Levine**, “Deviations from Optimum Information-Purchase Strategies in Human Decision-Making,” *Organizational Behavior And Human Performance*, July 1970, *5* (4), 313–329.
- Kahneman, Daniel and Amos Tversky**, “Subjective Probability: A Judgment of Representativeness,” *Cognitive Psychology*, July 1972, *3* (3), 430–454.
- and —, “On the psychology of prediction,” *Psychological Review*, 1973, *80* (4), 237–251.
- and **Dan Lovall**, “Timid Choices and Bold Forecasts: A Cognitive Perspective on Risk Taking,” *Management Science*, 1993, *39* (1), 17–31.
- Keren, Gideon**, “Calibration and probability judgments: Conceptual and methodological issues,” *Acta Psychologica*, 1991, *77* (3), 217–273.

- **and Willem A. Wagenaar**, “Violation of Utility Theory in Unique and Repeated Gambles,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 1987, *13* (3), 387–391.
- Klos, Alexander, Elke U. Weber, and Martin Weber**, “Investment decisions and time horizon: Risk perception and risk behavior in repeated gambles,” March 2005. Working Paper.
- Kraemer, Carlo and Martin Weber**, “How Do People Take into Account Weight, Strength, and Quality of Segregated vs. Aggregated Data? Experimental Evidence,” *The Journal of Risk and Uncertainty*, September 2004, *29* (2), 113–142.
- Marks, D.F. and J.K. Clarkson**, “An Explanation of Conservatism in the Bookbag-and-Pokerchips Situation,” *Acta Psychologica*, April 1972, *36* (2), 145–160.
- Martin, David W.**, “Data Conflict in a Multinomial Decision Task,” *Journal of Experimental Psychology*, 1969, *82* (1), 4–8.
- **and Charles F. Gettys**, “Feedback and Response Mode in Performing a Bayesian Decision Task,” *Journal of Applied Psychology*, 1969, *53* (5), 413–418.
- Nelson, Mark W., Robert Bloomfield, Jeffrey W. Hales, and Robert Libby**, “The Effect of Information Strength and Weight on Behavior in Financial Markets,” *Organizational Behavior and Human Decision Processes*, November 2001, *86* (2), 168–196.
- Nisbett, Richard E. and Lee Ross**, *Human inference: Strategies and shortcomings of social judgment*, Englewood Cliffs, NJ: Prentice Hall, 1980.
- Peterson, C. R. and L.D. Phillips**, “Revision of Continuous Subjective Probability Distributions,” *IEEE Transactions on Human Factors in Electronics*, March 1966, *HFE-7* (1), 19–22.
- Peterson, Cameron R. and Alan J. Miller**, “Sensitivity of Subjective Probability Revision,” *Journal of Experimental Psychology*, 1965, *70* (1), 117–121.
- **and Richard G. Swensson**, “Intuitive Statistical Inferences about Diffuse Hypotheses,” *Organizational Behavior And Human Performance*, February 1968, *3* (1), 1–11.
- **, Robert J. Schneider, and Alan J. Miller**, “Sample Size and The Revision of Subjective Probabilities,” *Journal of Experimental Psychology*, 1965, *69* (5), 522–527.
- **, Wesley M. DuCharme, and Ward Edwards**, “Sampling Distributions and Probability Revisions,” *Journal of Experimental Psychology*, 1968, *76* (2), 236–243.

- Phillips, Lawrence D. and Ward Edwards**, “Conservatism in a Simple Probability Inference Task,” *Journal of Experimental Psychology*, 1966, 72 (3), 346–354.
- Phillips, L.D., W.L. Hays, and W. Edwards**, “Conservatism in Complex Probabilistic Inference,” *IEEE Transactions on Human Factors in Electronics*, March 1966, HFE-7 (1), 7–18.
- Pitz, Gordon F.**, “Sample size, likelihood, and confidence in a decision,” *Psychonomic Science*, 1967, 8 (6), 257–258.
- , “Information Seeking When Available Information Is Limited,” *Journal of Experimental Psychology*, 1968, 76 (1), 25–34.
- **and Helen Reinhold Barrett**, “Information Purchase In A Decision Task Following The Presentation Of Free Information,” *Journal of Experimental Psychology*, 1969, 82 (3), 410–414.
- Rabin, Matthew**, “Diminishing Marginal Utility of Wealth Cannot Explain Risk Aversion,” in Daniel Kahneman and Amos Tversky, eds., *Choices, Values, and Frames*, Cambridge University Press, 2000, pp. 202–208.
- , “Risk Aversion and Expected-Utility Theory: A Calibration Theorem,” *Econometrica*, 2000, 68 (5), 1281–1292.
- , “Inference by Believers in the Law of Small Numbers,” *The Quarterly Journal of Economics*, August 2002, 117 (3), 775–816.
- **and Dmitri Vayanos**, “The Gambler’s and Hot-Hand Fallacies: Theory and Applications,” *Review of Economic Studies*, 2010, 77 (2), 730–778.
- **and Georg Weizsacker**, “Narrow Bracketing and Dominated Choices,” *American Economic Review*, 2009, 99 (4), 1508–1543.
- **and Richard H. Thaler**, “Anomalies: Risk Aversion,” *The Journal of Economic Perspectives*, 2001, 15 (1), 219–232.
- Read, Daniel, George Loewenstein, and Matthew Rabin**, “Choice Bracketing,” *Journal of Risk and Uncertainty*, 1999, 19 (1-3), 171–197.
- Redelmeier, Donald A. and Amos Tversky**, “On the Framing of Multiple Prospects,” *Psychological Science*, 1992, 3 (3), 191–193.
- Samuelson, Paul A.**, “Risk and Uncertainty: A Fallacy of Large Numbers,” *Scientia*, 1963, 98, 108–113.

- Sanders, A.F.**, “Choice Among Bets and Revision of Opinion,” *Acta Psychologica*, 1968, *28*, 76–83.
- **and W. Ter Linden**, “Decision Making During Paced Arrival of Probabilistic Information,” *Acta Psychologica*, 1967, *27*, 170–177.
- Sasaki, Shunichiro and Toshiji Kawagoe**, “Belief Updating in Individual and Social Learning: A Field Experiment on the Internet,” May 2007. Working Paper.
- Shaked, M. and Y L. Tong**, “Comparison of experiments for a class of positively dependent random variables,” *Canadian Journal of Statistics*, 1990, *18*, 79–86.
- Shu, Suzanne and George Wu**, “Belief Bracketing: Can Partitioning Information Change Consumer Judgments?,” October 2003. Working paper.
- Smith, Lones and Peter Sorenson**, “Pathological Outcomes of Observational Learning,” *Econometrica*, 2000, *68* (2), 371–398.
- Strub, Michael H.**, “Experience and Prior Probability in a Complex Decision Task,” *Journal of Applied Psychology*, 1969, *53* (2), 112–117.
- Swart, J.H. De**, “Effects of Diagnosticity and Prior Odds on Conservatism in a Bookbag-And-Pokerchip Situation,” *Acta Psychologica*, February 1972, *36* (1), 16–31.
- Teigen, Karl Halvor**, “Subjective sampling distributions and the additivity of estimates,” *Scandinavian Journal of Psychology*, September 1974, *15* (1), 50–55.
- Tversky, Amos and Daniel Kahneman**, “Belief in the Law of Small Numbers,” *Psychological Bulletin*, 1971, *76* (2), 105–110.
- **and –**, “Rational Choice and the Framing of Decisions,” *The Journal of Business*, 1986, *59* (4), S251–S278.
- **and –**, “Loss Aversion in a Riskless Choice: A Reference-Dependent Model,” *The Quarterly Journal of Economics*, 1991, *106* (4), 1039–1061.
- **and Ward Edwards**, “Information Versus Rewards in Binary Choices,” *Journal of Experimental Psychology*, 1966, *71* (5), 680–683.
- Wald, Abraham**, *Sequential Analysis*, New York, NY: Wiley, 1947.

Wendt, Dirk, "Value of Information for Decisions," *Journal of Mathematical Psychology*, October 1969, *6* (3), 430–443.

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

Winkler, Robert L. and Allan H. Murphy, "Experiments in the Laboratory and the Real World," *Organizational Behavior And Human Performance*, October 1973, *10* (2), 252–270.