The Inference-Forecast Gap in Belief Updating*

Tony Q. Fan† Yucheng Liang‡ Cameron Peng§

April 27, 2023

Abstract

Evidence from experiments, surveys, and the field has uncovered both underreaction and overreaction to new information. We provide new experimental evidence on the underlying mechanisms of under- and overreaction by comparing how people make inferences *and* revise forecasts in the same information environment. Participants underreact to signals when inferring about underlying states, but overreact to the same signals when revising forecasts about future outcomes—a phenomenon we term "the inference-forecast gap." We show that this gap is largely driven by different simplifying heuristics used in the two tasks, and we provide evidence supporting both similarity and timing as plausible mechanisms.

*We thank Peter Andre, Nicholas Barberis, Daniel Benjamin, B. Douglas Bernheim, Stefano Cassella, Soo Hong Chew, Marcel Fafchamps, Cary Frydman, Nicola Gennaioli, Matthew Gentzkow, Thomas Graeber, Alex Imas, Jiacui Li, Shengwu Li, Chen Lian, Yueran Ma, Muriel Niederle, Ryan Oprea, Christopher Roth, Joshua Schwartzstein, Andrei Shleifer, Songfa Zhong, and audiences at various seminars and conferences for helpful comments. The RCT registry ID is AEARCTR-0007006. This study is approved by Stanford IRB in Protocol 44866, by CMU IRB in Protocol 2016_000000482, and by LSE Ethics Review (Ref: 23685). We are grateful for financial support from CMU, IZA, and LSE.

†Stanford University. Email: tonyqfan@stanford.edu.

‡Carnegie Mellon University. Email: ycliang@cmu.edu.

§London School of Economics and Political Science. Email: c.peng9@lse.ac.uk.

1 Introduction

When new information arrives, rational agents should update their beliefs according to Bayes' rule. Empirical research, however, has uncovered many instances in which agents' reactions to information deviate from Bayes' rule. One recurring theme in the study of belief updating is that the direction of belief-updating biases appears to vary from setting to setting. For example, one literature shows, in both the field and the lab, that individuals tend to *overreact* to recent news when asked to make *forecasts* (e.g., Hey, 1994; Greenwood and Shleifer, 2014; Gennaioli et al., 2016; Frydman and Nave, 2017; Conlon et al., 2018; Bordalo et al., 2020; Afrouzi et al., 2023). The concept of overreaction, in turn, has been used to explain anomalies such as excess volatility in financial markets and boom-bust cycles in the macroeconomy (e.g., Barberis et al., 2015; Maxted, 2020; Bordalo et al., 2021). However, another experimental literature shows, rather robustly, that when asked to make *inferences* about underlying states, participants typically underreact to realized signals (Benjamin, 2019). The notion of underreaction has been similarly cited to account for facts such as post-earnings announcement drifts in financial markets and households' sluggish responses to macroeconomic conditions (Barberis et al., 1998; Coibion and Gorodnichenko, 2015). Indeed, both overreaction and underreaction are key concepts in economic analysis and have spurred the development of theories tackling important puzzles in finance and macroeconomics. However, so far we still know little about what makes people overreact in some environments but underreact in others (Benjamin, 2019).

In this paper, by running a series of online experiments, we propose one key condition that mediates under- and overreaction to new information. The experiment is motivated by an apparent tension between the two aforementioned literatures that directly test Bayesian updating using reported beliefs. While this tension could be attributed to differences in contexts or data-generating processes (DGPs), we propose an alternative, unexplored explanation: belief updating differs between an inference problem and a forecast-revision problem. The differences between the two problems are illustrated in Figure 1. Loosely, an inference problem is one where an agent observes signals and learns about the underlying state that determines the distribution of signals. By contrast, a forecast-revision problem is one where an agent also observes signals but instead update beliefs about future outcomes whose distributions also depend on the underlying state.

In rational models and most behavioral ones, the forecast-revision problem is closely tied to the inference problem: inference about the underlying state often serves as the first step or the primary input to revising forecasts about future outcomes. However, we uncover a disconnect between the two problems: when participants perform both types of updating tasks, they underreact to signals when making inferences but overreact when revising forecasts. We run additional treatments to study the underlying mechanisms driving this difference, and we discuss the potential connections

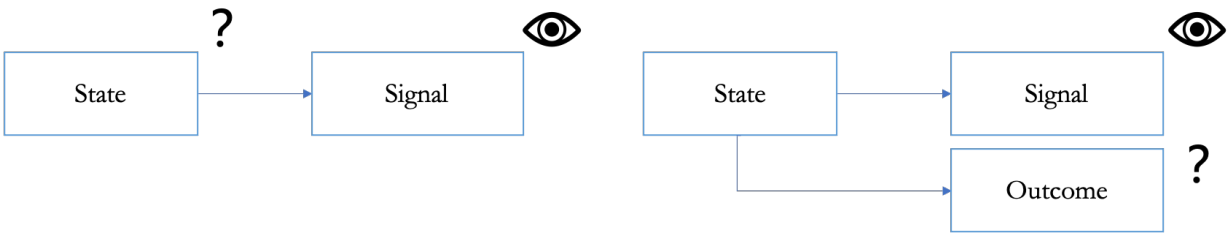

Figure 1: Inference problem (left) and forecast-revision problem (right)

Notes: In an inference problem, people observe a signal and then update their beliefs about the underlying states. In a forecast-revision problem, people revise their forecasts about outcomes in response to a realized signal.

of this result to non-experimental settings.

Our baseline treatment follows the "bookbag-and-poker-chip" paradigm in experimental research but frames the relevant variables in economic terms.[1] In each round of the experiment, there is a "firm" with a fixed state which is either good or bad. The firm generates signals, framed as its monthly stock price growth, and the signals are informative of the state; good firms, on average, have a higher growth in stock price than bad firms. Participants do not know the true state but are given the full DGP, including the prior distribution over the two states and the distributions of signals conditional on each state. In each month, the signal distribution is i.i.d. normal, with a mean of 100 if the state is good and 0 if it is bad.

The key to our design is to compare belief updating about underlying states and about future outcomes in the same information environment. There are two main parts in the baseline treatment: *Inference* and *Forecast Revision*. In *Inference*, participants observe one realized signal and report their updated beliefs about *the states*—the likelihoods of the firm being good and being bad. In *Forecast Revision*, participants also observe one realized signal, but instead report their updated expectations about *the next signal*—the expected stock price growth next month. In our environment, these two types of beliefs are closely linked: if one believes that the firm is good with a $p\%$ chance, then by the Law of Iterated Expectations (LoIE), the expectation about the next signal should be $p\% \times 100 + (1 - p\%) \times 0 = p$. This simple relationship ensures that, for participants who understand this link, the two problems involve a similar level of computational complexity.

Despite the straightforward connection between *Inference* and *Forecast Revision*, participants' behaviors and biases in the two tasks are qualitatively different. In *Inference*, 61% of the answers underreact relative to the Bayesian benchmark while 24% overreact, replicating the stylized fact of underreaction in the bookbag-and-poker-chip literature. By contrast, in *Forecast Revision*, 40%

[1] In a typical experiment under this paradigm, there is a bookbag that contains poker chips of several colors. Participants do not know the bag's color composition, but are given the prior distribution of the composition. A random chip is then drawn from the bag and, upon observing its color, participants are asked to report their posterior beliefs about the bag's color composition.

of the answers underreact while 54% overreact. Similarly, when belief updates are measured using the difference between posterior and prior beliefs, the average magnitude of belief updates is substantially larger for *Forecast Revision* than for *Inference*. We refer to this discrepancy in belief updating as the "inference-forecast gap." This gap is robust across subsamples, across rounds, and under alternative framings of the signal and the outcome. Moreover, the gap persists in two additional treatments: one in which the signal follows a binary distribution and one in which the outcome is different from the signal and completely determined by the state. These treatments not only demonstrate that the gap is robust to alternative DGPs, but also help rule out explanations based on, for example, misperceptions of signal autocorrelation and related phenomena such as the hot-hand bias.

After documenting the inference-forecast gap, we examine participants' decision procedures. The gap should not arise if, in *Forecast Revision*, participants correctly implement the standard "infer-then-LoIE procedure" by (a) first updating their beliefs about the states as in *Inference* and (b) then using these posterior beliefs to compute the expected value of the forecast outcome under the LoIE. One possibility is that participants intend to follow the infer-then-LoIE procedure, but make errors due to its complexity. We present multiple pieces of evidence against this possibility. In particular, we run a treatment that shows participants their own inference answers when they solve the corresponding forecast-revision problems, effectively reducing the two-step infer-then-LoIE procedure to a one-step procedure of simply applying the LoIE. The treatment, however, has little impact on the gap. Moreover, we confirm that participants are largely capable of applying the LoIE correctly when solving a standalone expectation-formation problem. These results suggest that, in general, participants are not using the infer-then-LoIE procedure in *Forecast Revision*—correctly or with errors. Instead, they resort to alternative nonstandard procedures.

What alternative decision procedures do participants use? We shed light on this question by detecting modal behaviors in the two updating tasks. In *Inference*, the modal behavior is "non-updates:" in 30% of the answers, the posterior equals the prior. In *Forecast Revision*, the fraction of non-updates drops to 22%; meanwhile, two other behaviors that rarely appear in *Inference* become modal. The first mode, representing 20% of the answers, is to answer 100 when the signal is good and 0 when it is bad. The answers mean that participants make forecasts as if they were 100% sure about being in the more representative state—the state more consistent with the signal—a simplifying heuristic that we term "exact representativeness." The second mode, constituting 12% of the answers, is to report a forecast that equals the signal itself. That is, participants directly use the realized signal as their expectation of the next outcome—a simplifying heuristic that we term "naive extrapolation." Each of the three modal behaviors corresponds to participants using a different salient cue in the information environment—the prior, the outcome expectation conditional on the representative state, and the realized signal—as an anchor in making forecasts (Kahneman

and Frederick, 2002; Shah and Oppenheimer, 2008). Moreover, excluding these modal behaviors would substantially reduce the inference-forecast gap, suggesting that they are largely responsible for the aggregate patterns.

What gives rise to these different simplifying heuristics? We propose and test two complementary mechanisms. First, we hypothesize that people are more likely to rely on salient cues that appear similar to the variable elicited by the belief-updating question (Slovic et al., 1990; Kahana, 2012; Bordalo et al., 2023). For example, the expected outcome conditional on the representative state is a salient cue in the information environment, and it appears more similar to the expected outcome conditional on the signal (elicited by *Forecast Revision*) than to the posterior probabilities of the states (elicited by *Inference*). Therefore, participants are more likely to anchor on this cue when revising forecasts, resulting in exact representativeness.[2] To test this hypothesis, we run two additional treatments that vary the similarity between informational cues and the belief-updating questions. In one treatment, for example, we increase the similarity between the inference variable and the two cues driving exact representativeness and naive extrapolation. The two heuristics become more prevalent among inference answers, reducing the inference-forecast gap.

Second, we explore whether the timing of the elicited variable plays a role. In all our treatments so far, the states are determined before the signals are realized while the forecast outcomes are realized in the future (i.e., after the signals). We hypothesize that the relative timing between the realization of states, signals, and outcomes may play a role in the higher prevalence of overreaction-inducing heuristics in *Forecast Revision* than in *Inference*, thus contributing to the inference-forecast gap. To test this hypothesis, we run an additional treatment in which we manipulate the relative timing between signal realization and outcome realization: participants update beliefs about outcomes in a *previous* month based on signals observed in the *current* month. The overreaction-inducing heuristics (especially naive extrapolation) become less prevalent among forecast revisions, again reducing the inference-forecast gap.

Our main results—namely, the inference-forecast gap and the use of heuristics—are based on an experimental setting that is simple and transparent. An important question remains as to how to apply these findings in more complex field settings. In the last part of the paper, we take a first stab at this question by (a) offering theoretical reasoning on how increased complexity in the field may strengthen the phenomena observed in our experiment, and (b) presenting suggestive evidence that the heuristics identified in our experiment also emerge in survey forecasts of real economic variables, among both professional forecasters and households. We also discuss the implications of our results for future research on belief formation, on both theoretical and empirical fronts.

[2]Analogously, the realized signal as a cue is more similar to the forecast-revision variable than to the inference variable, so it is more likely to serve as an anchor when participants revise forecasts, resulting in the behavioral mode of naive extrapolation.

Our work is related to an active body of experimental research seeking to understand the conditions of underreaction and overreaction in belief updating (He and Kucinskas, 2020; Enke et al., 2021; Hartzmark et al., 2021; Liang, 2022; Afrouzi et al., 2023; Enke and Graeber, 2023).[3] Recently, Ba et al. (2022) and Augenblick et al. (2023) also try to reconcile under- and overreaction in different settings. They focus on inference tasks under the bookbag-and-poker-chip paradigm to detect conditions that moderate under- and overreaction such as the number of states and the strength of signals. Unlike these papers, we fix the information environment and vary the type of belief-updating problems, and we find that underreaction in inference problems does not generalize to forecast-revision problems.[4] This brings a new perspective to this literature: the direction of belief-updating biases can depend on the types of belief elicited. Moreover, by connecting the inference-forecast gap to the use of different simplifying heuristics, we further highlight the role of complexity and incorrect mental models in explaining belief-updating biases (Enke and Zimmermann, 2019; Enke, 2020; Esponda et al., 2020; Andre et al., 2021; Graeber, 2021; Kendall and Oprea, 2022).

Not only do we document the inference-forecast gap, we also propose and provide evidence for two specific mechanisms driving the gap. First, we build on recent work on salience and memory retrieval (Kahana, 2012; Bordalo et al., 2023) and argue that the similarity between belief-updating questions and salient cues in the information environment plays an important role in determining the use of heuristics and the direction of belief-updating biases. Second, we find that beliefs about future outcomes react more to new signals than beliefs about past outcomes. This result contributes to a literature that shows how the timing of uncertainty affects decisions (Rothbart and Snyder, 1970; Heath and Tversky, 1991; Benjamin et al., 2017; Nielsen, 2020).[5]

We provide experimental evidence for overreaction in forecast-revision problems and discuss its implications for field settings. In this regard, our paper complements experimental studies on autocorrelated time-series forecasts (Hey, 1994; Frydman and Nave, 2017; He and Kucinskas, 2020;

[3]Empirical work using field or survey data, including Malmendier and Nagel (2011, 2016) and Wang (2020), also discusses the conditions under which people overreact and underreact to new information.

[4]A few belief-updating experiments using the bookbag-and-poker-chip design elicit beliefs of future draws conditional on the current draw. Moreno and Rosokha (2016), Bland and Rosokha (2021), Hartzmark et al. (2021) and Epstein and Halevy (2021) find either near-Bayesian updating or overreaction in their average results, and Fehrler et al. (2020) finds underreaction. None of these experiments compare beliefs of future draws with beliefs of the bookbag's composition.

[5]Our paper is also related to the psychology literature on the asymmetry between diagnostic reasoning ($\Pr(\text{Cause}|\text{Effect})$) and predictive reasoning ($\Pr(\text{Effect}|\text{Cause})$) in a given causal structure (e.g., Tversky and Kahneman, 1980; Fernbach et al., 2011). While the inference problem in our paper is synonymous to diagnostic reasoning, forecast revision is different from either kinds of reasoning in this literature because it elicits the belief of one "effect" (the forecast outcome) of the "cause" (the underlying state) conditional on another effect (the signal). Moreover, in parts of our experiments, we elicit forecasts without showing participants any signal, which is more akin to predictive reasoning. However, we show that biases in these parts cannot explain the inference-forecast gap. We thank Thomas Graeber for pointing us to this literature.

Afrouzi et al., 2023) to provide support for overreaction in survey expectations (e.g., Greenwood and Shleifer, 2014; Bordalo et al., 2020; Barrero, 2022). Unlike previous forecast experiments, DGPs in our experiment fully specify the underlying states, which in turn determine the signal and outcome distributions. This design brings the setting closer to standard models in macroeconomics and finance and lends several advantages to our analysis.[6] First, the explicit separation between states and outcomes makes it possible to design different problems targeting inference and forecast revision, respectively, thereby allowing us to pin down where a specific updating bias arises. Second, it allows us to separately identify the specific forms of overreaction, such as representativeness-based overreaction (Kahneman and Tversky, 1972; Bordalo et al., 2018) and mechanical extrapolation (Barberis et al., 2015, 2018). Third, having a fully-specified DGP allows us to attribute biases in posterior beliefs to incorrect statistical reasoning rather than to misperceived DGPs.

Overreaction in *Forecast Revision* is reminiscent of the hot-hand bias, the exaggeration of belief in an outcome after observing a long streak of the same outcomes (Gilovich et al., 1985; Tversky and Gilovich, 1989; Suetens et al., 2016).[7] In contrast, overreaction occurs in our experiment after just *one* signal realization. Moreover, we find overreaction even when the forecast outcome is different from the signal variable and fully determined by the state, a setting in which misperceptions of outcome autocorrelation and related phenomena such as the hot-hand bias, are irrelevant.[8] Overall, it is unlikely that our results are driven by the hot-hand bias.

The rest of the paper proceeds as follows. Section 2 outlines our experimental design. Section 3 documents the existence of the inference-forecast gap. Section 4 studies the decision procedures used by participants. Section 5 explores the mechanisms behind these decision procedures. Section 6 concludes and discusses the implications of our results.

[6] In asset-pricing models, when investors are learning about firm quality (fundamentals), it is common to assume that they observe noisy signals of quality such as stock returns (e.g., Glaeser and Nathanson, 2017). In the mutual fund literature, investors learn about manager skills by observing past fund returns (e.g., Berk and Green, 2004; Rabin and Vayanos, 2010). In the labor literature, job seekers learn about their employability from the offers they receive (Burdett and Vishwanath, 1988).

[7] The opposite phenomenon of the gambler's fallacy, which is more often observed in experiments (Benjamin, 2019), would predict more underreaction in forecast-revision tasks.

[8] Our underinference result is also inconsistent with the leading account of the hot-hand bias, which is based on overinference (Rabin, 2002; Rabin and Vayanos, 2010). At the design level, we use explicit instructions and comprehension checks to make sure participants do not commit the hot-hand bias.

2 Experimental Design

2.1 Environment

To compare belief updating between making inferences and revising forecasts for the same individual, we adopt a within-participant experimental design. For each inference problem a participant solves, there is a corresponding forecast-revision problem with the same information environment, i.e., the same DGP and the same realized signal.

The main treatment, *Baseline*, has five parts, summarized in Table 1. Each part has eight rounds of problems. In each round, participants are first presented with a "firm" randomly drawn from a new pool of 20 firms. A firm's state, θ, is either G(ood) or B(ad). Participants do not know the state of the drawn firm, but are given the composition of the pool, which specifies the prior distribution over the states. The firm generates signals, s_t, which are framed as the firm's stock price growth in month t. Participants are provided with the conditional distributions of signals: signals of a good firm follow an i.i.d. normal distribution of $N(100, \sigma^2)$ and signals of a bad firm follow i.i.d. $N(0, \sigma^2)$.[9] Because good firms are more likely to have higher stock price growth than bad firms, a signal of high stock price growth (higher than 50) is diagnostic of the firm being good.

Table 1: Summary of variables elicited in each part of *Baseline*

Number	Part	Show signal?	Beliefs elicited
1	*Inference Prior*	No	$\Pr(\theta)$
2	*Inference*	Yes	$\Pr(\theta \mid s_0)$
3	*Forecast Prior*	No	$\mathbb{E}(s_1)$
4	*Forecast Revision*	Yes	$\mathbb{E}(s_1 \mid s_0)$
5	*Expectation Formation*	No	$\mathbb{E}(s_1)$

To sum up, in each round, the DGP is fully specified by two pieces of information: the prior distribution of states and the conditional distributions of signals. Both are presented to participants using figures and texts in a one-page display (see Figure 2 for an example), and we explain this interface with detailed instructions.[10] Table 2 summarizes the parameter values for the eight DGPs. We include six DGPs with symmetric priors ($\Pr(G) = 50\%$) and two DGPs with asymmetric priors. The DGPs with symmetric priors allow us to identify underreaction and overreaction without confounds from base-rate neglect, while the DGPs with asymmetric priors help us examine the

[9] In the actual implementation, we discretize the supports of normal distributions to multiples of 10 and truncate at both tails.

[10] Screenshots of the experimental interface can be found in the Online Appendix.

robustness of our results. Each DGP is represented by one problem in each of the five parts (the DGP is modified in the *Expectation Formation* part, which we will explain later). As a result, answers across parts are directly comparable. Unless mentioned otherwise, an observation refers to a participant's answers to the five corresponding questions in all five parts.

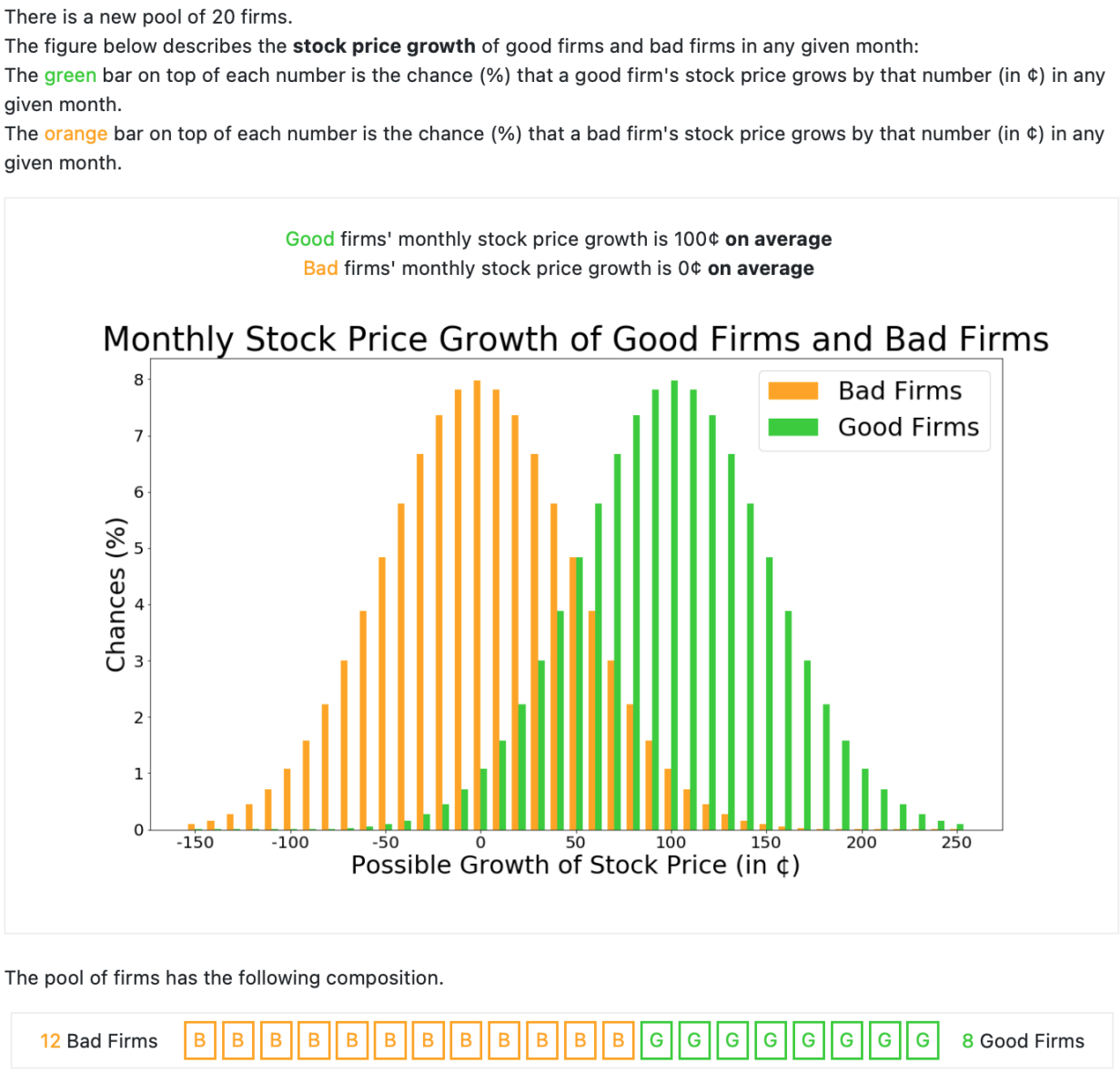

Figure 2: An example of the interface for the DGP

The two main parts of the experiment are *Inference* and *Forecast Revision*. For both parts, in each round, participants first observe the firm's stock price growth in the current month s_0. In *Inference*, after seeing the realized signal, participants report their updated beliefs about the states $\Pr(\theta|s_0)$. These beliefs are elicited in percentages, and henceforth we will refer to an inference

Table 2: Parameter values for DGPs

Index	1	2	3	4	5	6	7	8
$\Pr(G)$	50%	50%	50%	50%	50%	50%	80%	20%
σ	50	60	70	80	90	100	100	100

answer as the reported belief about the Good state without the % sign.[11] In *Forecast Revision*, participants instead report their updated expectations about the firm's stock price growth next month $\mathbb{E}(s_1|s_0)$. To allow for a direct comparison between the two parts, signal realization is set to be the same in any two corresponding rounds for the same participant, though it varies across participants.

In the other three parts, participants do not observe any signal realization before their beliefs are elicited. In *Inference Prior*, participants directly report prior beliefs about the states $\Pr(\theta)$ based on their knowledge about the DGP. Similarly, in *Forecast Prior*, they directly report prior expectations about the signal $\mathbb{E}(s_1)$. These two parts test whether participants can correctly form prior beliefs. The last part, *Expectation Formation*, is identical to *Forecast Prior*, except for the composition of firms in the pool. While the composition of firms in *Forecast Prior* is set exogenously according to Table 2, in *Expectation Formation* it is determined endogenously by participants' reported posterior beliefs about the states in *Inference*. For example, if a participant reports a posterior belief of $\Pr(G|s_0) = 40\%$ in a round in *Inference*, then the pool of firms in the corresponding round in *Expectation Formation* will have 8 ($= 40\% \times 20$) good firms and 12 bad ones.[12] *Expectation Formation* is designed to test whether participants can correctly form expectations about the next signal when the states are distributed according to their own inference posteriors.

To ensure that sufficient attention is paid to the problems and to prevent click-through behavior, participants need to stay on each page for at least eight seconds before they are allowed to type in their answers. For each participant, we further randomize (a) the order of different DGPs in each part and (b) the order of the five parts. For the latter randomization, we require that (a) priors are elicited before eliciting the corresponding posteriors and (b) the *Expectation Formation* part comes after *Inference*. Hence, we are left with three orders of parts: 12345, 12534, and 34125.

After the five parts, participants complete an unincentivized exit survey. At the end of the experiment, participants may receive a $5 bonus payment, and their chance of receiving the bonus

[11] In the experimental interface, there is one blank for the belief about the Good state and one for the Bad state. Once a participant types a number into one of the two blanks, the other blank will be automatically filled with 100 minus that number. Only numbers in the range $[0, 100]$ are allowed.

[12] The numbers of good and bad firms in *Expectation Formation* are rounded to the nearest integer if the reported beliefs in *Inference* are not a multiple of 5%. Fourteen percent of the answers in *Inference* are not multiples of 5%, among which half are rounded up and the other half rounded down.

depends on their answer in one randomly selected round through a quadratic rule.[13]

Building on *Baseline*, we implement several straightforward extensions as robustness checks. First, we frame the signal as revenue growth instead of stock price growth. Second, we ask participants about their expectations of the *last* signal s_{-1} ("stock price or revenue growth in the previous month") instead of the *next* signal s_1. In Appendix A.5, we show that results are qualitatively similar across all these extensions. Therefore, we pool the data from all versions of *Baseline* for our main results.

2.2 The no inference-forecast gap benchmark

According to standard probability theory, answers in *Inference* and *Forecast Revision* should be tightly linked. Specifically, the Law of Iterated Expectation (henceforth abbreviated as "LoIE") implies the following equation:

$$\mathbb{E}(s_1|s_0) = \Pr(G|s_0) \times \mathbb{E}(s_1|G, s_0) + \Pr(B|s_0) \times \mathbb{E}(s_1|B, s_0). \tag{1}$$

In our experiment, s_1 and s_0 are independent conditional on the state θ, so $\mathbb{E}(s_1|G, s_0) = \mathbb{E}(s_1|G) = 100$ and $\mathbb{E}(s_1|B, s_0) = \mathbb{E}(s_1|B) = 0$. Therefore, Equation (1) simplifies to the following equation:

$$\mathbb{E}(s_1|s_0) = \Pr(G|s_0) \times 100. \tag{2}$$

We term Equation (2) the "no inference-forecast gap" condition. It summarizes the theoretical link between the posterior belief about the underlying states and the updated forecast of the outcome variable s_1. If an inference answer and its corresponding forecast-revision answer satisfy this condition, then there should be no discrepancy between these two types of belief-updating problems: Bayesian inference would translate to rational forecasts, and any deviation from Bayes' rule in the inference answer would imply the same deviation from rationality in the forecast-revision answer.

The computational simplicity of Equation (2) is an advantage of our experimental design. Under the no inference-forecast gap condition, if a signal leads to a belief that the good state has 40% probability, then the resulting expectation of the outcome should be 40. For participants who understand this condition, the computational cost of solving a forecast-revision problem is very close to that of solving the corresponding inference problem. Therefore, computational complexity alone is unlikely to cause violations of the no inference-forecast gap condition.[14]

[13] If their answer in that round equals the rational benchmark according to standard probability theory, then they receive the bonus with certainty; otherwise, their chance of getting the bonus decreases quadratically in the difference between their answer and the rational benchmark (see (Hartzmark et al., 2021) for a similar incentive structure). If the answer is p and the rational benchmark is q (in % for the two *Inference* parts), then the chance of receiving the bonus is $\max\{0, (100 - (p - q)^2)\%\}$.

[14] Moreover, because beliefs are equally incentivized across the two types of problems, rational tradeoffs between

When participants solve a forecast-revision problem, one simple and standard procedure that satisfies the no inference-forecast gap condition is the following "infer-then-LoIE" procedure: In the first step, participants update their beliefs about the states using the same (and possibly non-Bayesian) rule as in the corresponding inference problem; in the second step, they apply the LoIE using the posteriors from the first step to obtain their expectations about the forecast outcome. In later parts of the paper, we will examine whether participants follow this procedure.

2.3 Instructions and comprehension questions

Participants receive extensive instructions, with the tasks and incentive structure explained in detailed and intuitive terms. In particular, we go to great lengths to ensure that participants fully understand the DGP. First, we emphasize that the state of a firm is constant across months but the signals are i.i.d. conditional on the state. In doing so, we explicitly caution against incorrect beliefs that the signals are autocorrelated conditional on the state. Second, we use an example DGP to illustrate the discretized normal distributions of the signals. In particular, we highlight the conditional means (0 and 100) and the property that signals higher (lower) than 50 are good (bad) news about the firm's quality. Third, we present participants with two explicit formulae, one for calculating the prior distribution over states from the pool composition ($\Pr(G) = \frac{\text{number of Good firms}}{20}$) and one for calculating the expectation about the signal from the belief about the states ($\mathbb{E}(s) = \Pr(G) \times 100$). However, we do not mention or nudge participants toward any specific belief-updating rule.

At the end of the instructions, participants need to answer a set of comprehension questions that test their understanding of the DGP, the incentive structure, and the two formulae. Participants can proceed only if they have answered all the comprehension questions correctly.[15]

2.4 Procedural details

We programmed our experiment using oTree (Chen et al., 2016). For *Baseline*, we recruited 279 participants through Prolific, an online platform designed for social science research.[16] Signals were framed as monthly revenue growth for 142 participants and as stock price growth for 137 participants. There was also some variation across participants in the order of parts: 102 participants went through the experiment in the order of 12345, 103 in the order of 12534, and 74 in

monetary gains and computational costs, in the spirit of Sims (2003); Gabaix (2014); Caplin and Dean (2015); and Woodford (2020), cannot generate an inference-forecast gap.

[15] If there are mistakes, participants will be asked to re-answer those questions.

[16] See Palan and Schitter (2018) on using Prolific as a participant pool. We recruited only US participants who had completed more than 100 tasks on Prolific and who had an approval rate of at least 99%.

the order of 34125. The participants, on average, spent about 30 minutes on the experiment and earned a payment of $7.08, $5 of which was the base payment.

2.5 Other treatments

In addition to *Baseline*, we also implemented several other treatments to investigate the robustness of and the mechanisms behind our results. These treatments are summarized in Table 3. Details about these treatments will be described in their respective sections.

Table 3: Overview of additional treatments

Treatment	Section	Key differences from *Baseline*
Deterministic Outcome	3.2	Forecast outcome is a different variable (100 if $\theta = G$ and 0 if $\theta = B$)
Binary Signal	3.3	Signals are binary; forecast questions ask about full distributions
Nudge	4.1	Beliefs about states and forecasts are elicited on the same page
More Similar	5.1.2	State variable (profitability) = mean of signal or forecast outcome (profits); inference questions ask about the expectation of the state
Less Similar	5.1.3	Forecast outcome is a different variable (up if $\theta = G$ and down if $\theta = B$); forecast questions ask about full distributions

3 Evidence for the Inference-Forecast Gap

3.1 Aggregate patterns

In this section, we compare belief updating between inference and forecast-revision problems using two methods of analysis. First, we classify each answer into one of three categories—Near-rational, Underreaction, and Overreaction—and examine the distributions of answers by categories. Second, we calculate the average belief movement from the prior to the posterior. Recall that, if the no inference-forecast gap condition in Equation (2) is met, then results from *Inference* and *Forecast Revision* should exhibit similar patterns. Any systematic difference, therefore, would imply an inference-forecast gap.

For an inference problem in our experiment, the rational benchmark is given by Bayes' rule:

$$\Pr^{\text{Rational}}(G|s_0) = \frac{\Pr(G) \cdot \Pr(s_0|G)}{\Pr(G) \cdot \Pr(s_0|G) + \Pr(B) \cdot \Pr(s_0|B)}. \tag{3}$$

For a forecast-revision problem in our experiment, the rational benchmark can be derived by applying LIE to the corresponding rational inference answer:

$$\begin{aligned}\mathbb{E}^{\text{Rational}}(s_1|s_0) &= \Pr^{\text{Rational}}(G|s_0) \times \mathbb{E}(s_1|G) + \Pr^{\text{Rational}}(B|s_0) \times \mathbb{E}(s_1|B) \\ &= \Pr^{\text{Rational}}(G|s_0) \times 100. \end{aligned} \tag{4}$$

Note that the no inference-forecast gap condition in Equation (2) is satisfied by the rational benchmarks.

We first classify answers in *Inference* and *Forecast Revision* by how they compare to the rational benchmarks. An answer is classified as Near-rational if its difference from the rational benchmark is no more than 2.5.[17] To introduce the categories of Underreaction and Overreaction, we first define an "update" by how much an answer moves from its (objective) prior value in the direction of the realized signal s_0:

$$\text{update} = \begin{cases} \text{answer} - \text{prior}, \text{ if } s_0 > 50 \\ \text{prior} - \text{answer}, \text{ if } s_0 < 50 \end{cases}. \tag{5}$$

For any two corresponding inference and forecast-revision problems, Equations (3) and (4) imply that their rational updates are identical. We classify an answer as Underreaction (Overreaction) if the update is smaller (larger) than the rational update by more than 2.5; we do not classify answers when $s_0 = 50$, i.e., the realized signal is uninformative.

Table 4 shows the aggregate patterns in *Baseline* (excluding observations with a signal of 50). The first three columns concern the distribution of answers by categories. Results from *Inference* replicate the key finding from the classic bookbag-and-poker-chip literature: participants overwhelmingly underreact to new information and update too little about the firm's underlying state. Out of all the answers, 60.8% are Underreaction, 24.1% are Overreaction, and 15.2% are Near-rational. These patterns, however, flip in *Forecast Revision*: 53.9% of the answers indicate overreaction to new information, higher than the fraction of 39.7% classified as Underreaction.

The last column of Table 4 concerns the average update. In *Inference*, the average update is 14.3, significantly lower than the average rational update of 23.3 ($p < 0.01$). By contrast, in *Forecast Revision*, the average update is 32.7, significantly higher than the rational benchmark ($p < 0.01$). Therefore, both methods of analysis suggest an inference-forecast gap. In the Appendix, Table A6 further confirms the statistical significance of the inference-forecast gap in a regression framework.

The inference-forecast gap is highly robust in various cuts of the data (see Section A of the

[17] We choose the number 2.5 so that the interval for near-rational covers at least one multiple of five, on which participants' answers tend to cluster.

Table 4: Aggregate patterns in *Baseline*

N=279, Obs.=2144	Classification			Update
	Underreaction	Near-rational	Overreaction	Mean (s.e.)
Inference	60.8%	15.2%	24.1%	14.3 (.7)
Forecast Revision	39.7%	6.4%	53.9%	32.7 (2)
Rational				23.3 (0.3)

Notes: The first three columns present the percentages of answers classified as Underreaction, Near-rational, and Overreaction. The last column shows the average belief movement from the (objective) prior to the posterior, as well as the rational benchmark. Observations with the signal equal to 50 are excluded. Standard errors are clustered by participant.

Appendix for details). First, in a more "reasonable" subsample that only includes observations with (a) answers that fall within $[0, 100]$ and (b) updates in the correct direction, *Forecast Revision* no longer exhibits overreaction on average, but the inference-forecast gap remains highly significant. Second, the gap is present under all eight DGPs, even though they entail different priors and signal distributions. Third, the gap increases for stronger signals—that is, when the signal deviates more from 50 and therefore becomes more informative—but exists even for the weakest signals. Fourth, our results persist in a subsample that excludes observations with incorrect reported prior beliefs. Fifth, there is no qualitative impacts on the inference-forecast gap (a) when we change the order of experimental parts, (b) when the signal and outcome are framed as revenue growth, and (c) when we control for participant characteristics.

3.2 *Deterministic Outcome* treatment

In this and the next subsection, we investigate the inference-forecast gap in two additional treatments with alternative DGPs. In *Baseline*, the forecast outcome and the realized signal are part of the same time series. Therefore, the observed inference-forecast gap could be due to misperceived signal autocorrelation and related phenomena such as the hot-hand bias (Gilovich et al., 1985; Tversky and Gilovich, 1989; Suetens et al., 2016). To rule out this explanation, we implement an additional treatment called *Deterministic Outcome*.

In this treatment, the outcome variable in *Forecast Revision* is different from the signal variable: when the outcome variable is the firm's stock price growth, the signal variable is the revenue growth, and vice versa. Moreover, the outcome variable is fully determined by the state: it equals 100 for sure in the Good state and 0 for sure in the Bad state. The distributions of the state and the signal are the same as in *Baseline*. Under this alternative DGP, the no inference-forecast

gap condition remains the same: the forecast-revision answer equals the corresponding inference answer (minus the % sign). But unlike in *Baseline*, the perceived correlation between the signal and the outcome should be irrelevant for the inference-forecast gap here: since the outcome is fully determined by the state, the perceived signal-outcome correlation should be the same as the perceived signal-state correlation.

Table 5 shows a similar inference-forecast gap for *Deterministic Outcome* compared to *Baseline*. In the Appendix, Table A10 further confirms, in a regression analysis, that the gap is statistically significant.

Table 5: Aggregate patterns in *Deterministic Outcome*

N=100, Obs.=777	Classification			Update
	Underreaction	Near-rational	Overreaction	Mean (s.e.)
Inference	64.4%	14.8%	20.8%	13.4 (1.3)
Forecast Revision	39.9%	8.6%	51.5%	34.1 (3.5)
Rational				23.1 (.4)

Notes: The first three columns present the percentages of answers classified as Underreaction, Near-rational, and Overreaction. The last column shows average belief movements in the signal direction from the (objective) priors and their rational benchmark. Observations with the signal equal to 50 are excluded. Standard errors are clustered by participant.

Results from *Deterministic Outcome* clearly show that the hot-hand bias cannot account for the inference-forecast gap. This further differentiates our results from overreaction in univariate forecasts (Hey, 1994; Frydman and Nave, 2017; Afrouzi et al., 2023) in which exaggerated autocorrelation is a key driving force. Moreover, the treatment helps address two additional robustness issues. First, the inference-forecast gap is not limited to cases where the signal and the outcome share the same variable name and distribution. Second, even when the state variable and the outcome variable share the same distribution, an inference-forecast gap can still arise.

3.3 *Binary Signal* treatment

In a second treatment called *Binary Signal*, the signal s_t follows a binary distribution instead of a continuous distribution. In particular, the signal is framed as the direction of the firm's stock price movement and is either up or down, and the probability of an upward movement is higher if the firm's state is Good. The parameters for the DGPs are listed in Table 6. In the *Forecast Revision* part of this treatment, the problem asks about the full probability distribution of the outcome $\Pr(s_1)$, instead of the expectation $\mathbb{E}(s_1)$.

Table 6: Parameter values for DGPs in *Binary Signal*

Index	1	2	3	4	5	6	7	8
$\Pr(G)$	50%	50%	50%	50%	50%	50%	80%	20%
$\Pr(\text{up}\|G)$	60%	70%	80%	90%	70%	55%	70%	70%
$\Pr(\text{up}\|B)$	40%	30%	20%	10%	45%	30%	30%	30%

As in *Baseline*, the no inference-forecast gap condition in *Binary Signal* is given by the LIE:

$$\Pr(s_1 = \text{up}|s_0) = \Pr(G|s_0) \times \Pr(\text{up}|G) + \Pr(B|s_0) \times \Pr(\text{up}|B). \tag{6}$$

Substituting in $\Pr(\text{up}) = \Pr(\text{up}|G) \times \Pr(G) + \Pr(\text{up}|B) \times \Pr(B)$, which is the LIE applied to the objective prior beliefs, we obtain the following equation:

$$\frac{\Pr(s_1 = \text{up}|s_0) - \Pr(\text{up})}{\Pr(\text{up}|G) - \Pr(\text{up}|B)} = \Pr(G|s_0) - \Pr(G). \tag{7}$$

Equation (7) states that under the no inference-forecast gap condition, the inference update equals the *normalized* forecast-revision update, defined by how much the forecast revision answer moves from the objective prior in the signal direction *divided by* the range of outcome probabilities, $\Pr(\text{up}|G) - \Pr(\text{up}|B)$. This equation is not as simple as Equation (2) in *Baseline*, so computational complexity could confound the comparison between inference and forecast revision answers.[18] However, one advantage of the *Binary Signal* treatment is that it is closer to the common design in the bookbag-and-poker-chip paradigm (Benjamin, 2019).

In *Binary Signal*, the three categories—Near-rational, Underreaction, and Overreaction—are defined in the same way as in *Baseline*, except that the categories for forecast-revision answers are defined based on their *normalized* updates. Table 7 reports the results from *Binary Signal*. As in *Baseline*, more answers are classified as Overreaction in *Forecast Revision* than in *Inference*, and the average update in the former part is also larger.[19] However, answers in *Forecast Revision* do not exhibit overreaction on average. Overall, the *Binary Signal* treatment shows that the inference-forecast gap extends to environments with alternative signal distributions. It also shows that this phenomenon can persist when the elicited object in *Forecast Revision* is the full distribution of the outcome instead of its expected value.

[18]For example, computational complexity could lead to higher degrees of cognitive uncertainty (Enke and Graeber, 2023). This could push forecast-revision answers toward underreaction.

[19]In the Appendix, Table A11 shows in a regression that the gap in updates is significant at the 10% level.

Table 7: Aggregate patterns in *Binary Signal*

N=140, Obs.=1120	Classification			Update
	Underreaction	Near-rational	Overreaction	Mean (s.e.)
Inference	61.0%	20.1%	18.9%	11.0 (0.9)
Forecast Revision	54.9%	6.7%	38.4%	14.2 (2.2)
Rational				18.7 (0.0)

Notes: The first three columns present the percentages of answers classified as Underreaction, Near-rational, and Overreaction. The last column shows average belief movements in the signal direction from the (objective) priors and their rational benchmark. The updates of forecast-revision answers are normalized by $Pr(\text{up}|G) - Pr(\text{up}|B)$ so that they are comparable to the inference updates. Standard errors are clustered by participant.

4 Decision Procedures

To investigate the mechanisms driving the inference-forecast gap, we next examine the decision procedures used by participants in forecast-revision problems. As discussed in Section 2.2, the inference-forecast gap should not arise if participants correctly implement the infer-then-LoIE procedure by: (i) first updating their beliefs about the states, in the same way as in the inference problems, and (ii) then applying the LoIE to form expectations about the forecast outcome. The existence of an inference-forecast gap thus rejects that participants correctly implement this procedure in *Forecast Revision*. However, it is still possible that participants simply implement this procedure *incorrectly*: that is, they *intend* to follow the infer-then-LoIE procedure, but make errors because of the two-step nature of the procedure. In Section 4.1, we argue that this is unlikely to be the case. Then, in Section 4.2, we analyze what alternative procedures participants use.

4.1 Implementation errors or alternative procedures?

In this section, we present three pieces of evidence that go against the hypothesis that participants intend to follow the infer-then-LoIE procedure but simply make errors when they implement this procedure. In summary, we find that: (a) a treatment that reduces the complexity of the procedure does not significantly reduce the inference-forecast gap; (b) there is a very weak correlation between underreaction (overreaction) in inference problems and underreaction (overreaction) in forecast-revision problems; and (c) participants react significantly to the prior variance of the signal in inference problems but not in forecast-revision problems. Next, we detail these results in succession.

4.1.1 Reducing the complexity of the infer-then-LoIE procedure

If the two-step nature of the infer-then-LoIE procedure causes participants to make errors in implementing this procedure, then reducing the complexity of the procedure should mitigate such errors and reduce the inference-forecast gap. To test this hypothesis, we run an additional treatment, *Nudge*: in experimental parts that provide signals, after observing the realized signal, participants are first asked to report their beliefs about the states; and then, while the answers they just typed in are still on the screen, they are asked to report their expectations about the next signal.[20] For a participant intending to follow the infer-then-LoIE procedure, this design makes a forecast-revision problem no more complex than simply applying the LoIE: one only needs to multiply the inference posterior by 100 to complete the infer-then-LoIE procedure. In fact, because the inference question is quoted in percentage terms and the forecast-revision question in cents, participants can just type in the exact same number.

However, we find that displaying their own inference answers when participants revise their forecasts does not change the overall pattern of the inference-forecast gap. Table 8 shows the aggregate patterns in *Nudge*. Same as in *Baseline*, participants overwhelmingly underreact in *Inference* and on average overreact in *Forecast Revision*.[21]

How can we explain the persistence of the inference-forecast gap in *Nudge*? One possibility is that while the treatment indeed makes the infer-then-LIE procedure no more complex than solving a standalone expectation-formation problem, even the latter is error-prone for our participants, and the errors lead to overreaction. To test this possibility, in another part of *Nudge* called *Expectation Formation*, we ask participants to solve a standalone expectation-formation problem *without* seeing any signal realization. Specifically, in each round, we set the distribution over states in the expectation-formation problem to match the participant's own posterior beliefs reported in the corresponding inference problem. For example, if a participant reports $\Pr(G|s_0) = 40\%$ in a round in *Inference*, then the pool of firms in the corresponding *Expectation Formation* round will have 8 ($= 40\% \times 20$) good firms and 12 bad ones.[22]

Figure 3 plots the average deviation from LoIE in expectation-formation problems by the prior (probability of the Good state) and shows that, on average, the deviation is small in magnitude across the board. Moreover, in the third row of Table 8, we classify expectation-formation answers and calculate their updates.[23] Comparing the average update in *Inference*, *Forecast Revision*, and

[20] More specifically, participants have to stay on the page for eight seconds before answering each question. The forecast-revision question appears only after the answer to the inference question has been submitted. Participants can revise their answers to the inference question before they submit their answers to the forecast-revision question. The design of this treatment is similar to the *Nudge* treatments in Enke (2020) in spirit.

[21] In fact, the inference-forecast gap in *Nudge* is even larger than in *Baseline*, according to the regression analysis in Table A10.

[22] We implement a similar part in *Baseline* as well, and the results are similar (see Section C in the Appendix).

[23] Similar to before, the update of an expectation-formation answer is defined as the answer minus the (objective)

Table 8: Aggregate patterns in *Nudge*

N=100, Obs.=750	Classification			Update
	Underreaction	Near-rational	Overreaction	Mean (s.e.)
Inference	70.9%	10.0%	19.1%	10.1 (1.3)
Forecast Revision	41.3%	6.4%	52.3%	29.8 (3.0)
Expectation Formation	60.0%	6.7%	33.3%	14.7 (2.3)
Rational				22.5 (.5)

Notes: The first three columns present the percentages of answers classified as Underreaction, Near-rational, and Overreaction. The last column shows average belief movements in the signal direction from the (objective) priors and their rational benchmark. The expectation-formation answers are analyzed in the same way as the corresponding forecast-revision answers: the update of an expectation-formation answer is defined as the answer minus the (objective) prior in the corresponding forecast-revision problem if the signal in the latter problem is greater than 50 and the reverse if the signal is smaller than 50. The classification of an expectation-formation answer is conducted against the rational benchmark for the corresponding forecast-revision problem. Observations with the signal equal to 50 are excluded. Standard errors are clustered by participant.

Expectation Formation, we find that mistakes in *Expectation Formation* can account for only 23% ($= \frac{14.7-10.3}{29.8-10.3}$) of the inference-forecast gap. Therefore, it is unlikely that the inference-forecast gap stems from the mistakes participants make in standalone expectation-formation problems. All in all, results from *Nudge* suggest that the inference-forecast gap does not seem to result from complexity-induced errors.

4.1.2 Correlation between updating biases in inference and forecast revision

If participants generally follow the infer-then-LoIE procedure in *Forecast Revision*, then we should expect updating biases in *Inference* to be highly correlated with those in *Forecast Revision*. However, we find that updating biases in *Inference* are only weakly correlated with biases in *Forecast Revision*. For example, at the problem level, the correlation between overreaction in inference problems and overreaction in forecast-revision problems is only 0.07. At the participant level, the correlation between the fraction of overreactions in *Inference* and the fraction of overreactions in *Forecast Revision* is only 0.12.[24] We find very similar results when we study underreaction instead of overreaction in each case. This weak correlation further casts doubt on the possibility that

prior in the corresponding forecast-revision problem if the signal in the latter problem is greater than 50 and the reverse if the signal is smaller than 50. The classification of an expectation-formation answer is conducted against the rational benchmark for the corresponding forecast-revision problem.

[24]These correlations do increase in the *Nudge* treatment, to 0.27 and 0.25 respectively, which is to be expected. However, they are still far from perfect.

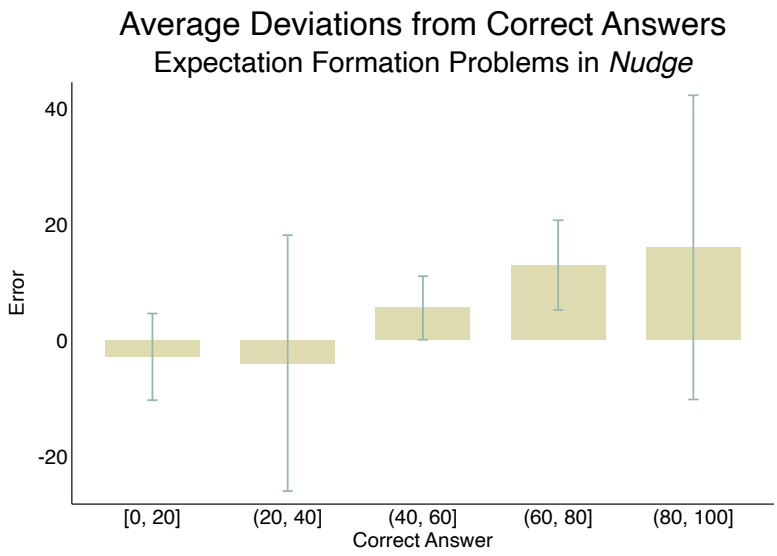

Figure 3: Deviations from LIE in expectation-formation problems by prior

Notes: We divide the expectation-formation problems in *Nudge* into five groups depending on the priors, and calculate the average error (deviation from correct answers) for problems in each group. Standard errors are clustered by participant.

participants follow the infer-then-LoIE procedure.

4.1.3 Reaction to the prior variance of the signal in updating

Recall that we vary the prior variance of the signal among the problems with symmetric priors (see Table 2). We exploit this feature of our design by testing whether participants respond to the standard deviation of the signal in updating, separately for *Inference* and *Forecast Revision*. Specifically, we use the following linear specification:[25]

$$\text{Absolute Update} = \beta \cdot \text{Signal Conditional SD} + \text{Signal Value FE} + \text{Participant FE} + \epsilon \qquad (8)$$

This regression essentially tests whether participants update less to signals of a given value (e.g., 90) when the conditional standard deviation of the signals is larger, as a Bayesian agent would do.

We estimate this equation separately for Bayesian updates, *Inference* updates, and *Forecast Revision* updates, and report the results in Table 9. Column (1) simply confirms that a Bayesian agent updates less to a given signal when the signal's conditional standard deviation is higher. Column (2) shows a similar pattern in *Inference* that is smaller in magnitude, indicating that participants indeed react to signal variance but are less sensitive than what Bayesianism implies. Column

[25]The relationship between signal standard deviation and the Bayesian update is not exactly linear, and also varies with the value of the signal. For ease of presentation, we adopt the linear specification as a reasonable approximation.

(3) shows that, in *Forecast Revision*, the reaction to signal variance is small and statistically insignificant. If participants actively use their inferences as input when they revise their forecasts, we should expect this coefficient to be much larger in magnitude and closer to the coefficient in Column (2).

Table 9: Does the amount of update respond to signal standard deviation?

	Absolute Update: |Posterior − Prior|		
	Bayesian	*Inference*	*Forecast Revision*
	(1)	(2)	(3)
Signal Conditional SD (50 $\sim$ 100)	-0.434***	-0.163***	-0.016
	(0.005)	(0.024)	(0.056)
Signal Value FE	Yes	Yes	Yes
Participant FE	Yes	Yes	Yes
Observations	1604	1604	1604
R^2	0.974	0.616	0.607

Notes: *, **, and *** indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively. We only use problems with a prior probability of 50% for the Good state; further, observations with the signal equal to 50 are excluded. Standard errors are clustered by participant.

Taken together, results in this section suggest that participants do not appear to be following the infer-then-LoIE procedure when solving forecast-revision problems—correctly or with errors. Rather, they appear to be using alternative procedures.

4.2 Alternative decision procedures

What alternative decision procedures do participants use in *Forecast Revision*? To answer this question, we examine the distributions of posterior beliefs to detect potential modal behaviors. To illustrate, Figure 4 plots the answer against the realized signal for problems with symmetric objective priors in *Inference* and *Forecast Revision*.[26] In *Inference*, a large fraction of answers equals the 50-50 prior, suggesting that many participants do not update based on the signal. The prevalence of such non-updates replicates a stylized fact in previous inference experiments (e.g., Coutts, 2019; Graeber, 2021).

[26]Distributions of answers in problems with asymmetric priors display similar patterns. See Appendix B for details.

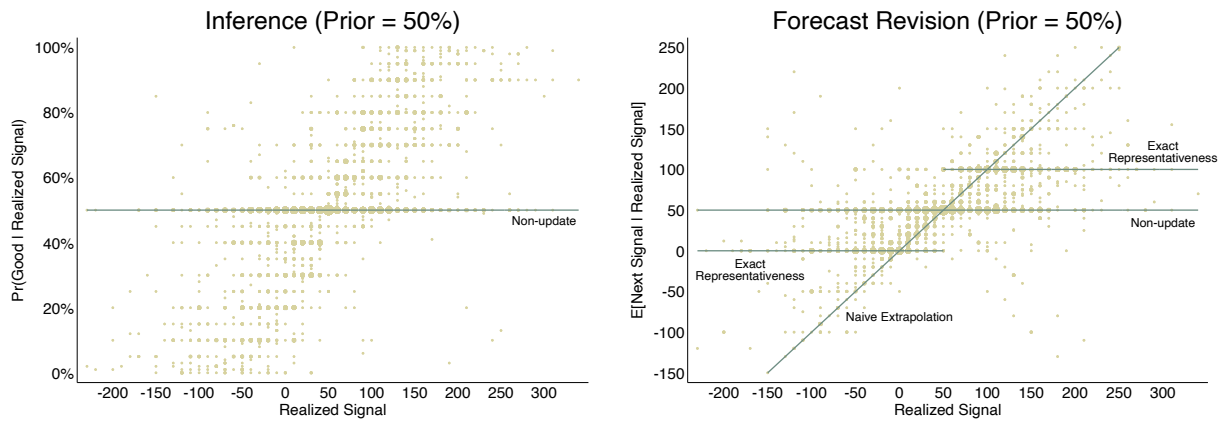

Figure 4: Scatterplots of answers against realized signals: subsample with symmetric priors

Notes: This figure plots the updated beliefs against the realized signals. The size of each circle represents the number of answers that equal the value on the y-axis given the realized signal on the x-axis. We only use problems with a prior probability of 50% for the Good state. In the right panel (the figure for *Forecast Revision*), we limit the range of the y-axis to [-150, 250] and drop observations that fall outside this range.

For *Forecast Revision*, non-updates also constitute a mode, shown by a cluster of answers that equal the 50-50 prior. However, two other modes also emerge. First, many forecast-revision answers cluster at 100 when $s_0 > 50$ and at 0 when $s_0 < 50$. Participants who give these answers behave as if they were certain about being in the representative state (the state consistent with the direction of the signal realization) and base their forecasts solely on that state. We term this overreacting behavior "exact representativeness" because it is consistent with the representativeness heuristic (Kahneman and Tversky, 1972; Bordalo et al., 2018).[27] This behavior is also consistent with a type of belief-updating process induced by coarse thinking (Mullainathan et al., 2008). Specifically, when updating beliefs, people consider only a finite set of categories rather than the full continuum of categories, and they change categories only when they see enough data to suggest that an alternative category better fit the data (Mullainathan, 2002).

Second, a smaller yet still significant fraction of forecast-revision answers are anchored at the face value of the realized signal.[28] We term this behavior "naive extrapolation" because it suggests a particular form of extrapolative beliefs whereby participants directly (and naively) use the most recent realization as their forecast for the next realization (Barberis et al., 2015, 2018;

[27] Note that our notion of exact representativeness is different from that in Camerer (1987), who first introduced the term.

[28] For each x-axis value—that is the value of the realized signal—we rank answers by the frequency of their occurrence. For 19 out of the 53 x-axis values, anchoring on the signal value is among the top three most frequent answers. In comparison, non-updates and exact representativeness are each among the top two most frequent answers for 36 x-axis values.

Liao et al., 2021).[29] This behavior leads to overreaction in the problems with symmetric priors in our experiment.

In Table 10, we define the behavioral modes and quantify their prevalence in *Baseline*. Confirming the patterns in the scatterplots, non-updates are widespread in both *Inference* and *Forecast Revision*, making up 29.7% and 21.9% of all answers, respectively. The other two behavioral modes, exact representativeness and naive extrapolation, appear almost exclusively in *Forecast Revision*, making up 20.3% and 11.9% of the answers, respectively. Only 3.3% of the answers meet the no inference-forecast gap condition and are not in any of the three behavioral modes. We conduct further analysis in Appendix B, where we find robust results when we relax the classification criteria for the modes and when we classify the participants rather than the answers.[30] At the participant level, we also document a modest degree of consistency between a participant's types in the two parts. For example, many participants are classified as non-updaters in both parts. We also present results on the modal behaviors in three other treatments, *Deterministic Outcome*, *Binary Signal*, and *Nudge*, and we find similar patterns.

Table 10: Modes of behavior in *Baseline*

Mode	Criterion for answer	*Inference*	*Forecast Revision*
Non-update	$=$ prior	29.7%	21.9%
Exact representativeness	$= 100$ if $s_0 > 50$, $= 0$ if $s_0 < 50$	2.6%	20.3%
Naive extrapolation	$= s_0$	3.2%	11.9%
No inference-forecast gap (excluding the other modes)	inference $=$ forecast revision	3.3%	
Unclassified		61.8%	45.2%
Observations		2144	2144

Notes: The column "Criterion for answer" shows the criterion for an answer to be classified into a mode. Note that an answer may be classified into more than one mode. The percentages in the last two columns are the fractions of answers in each mode in *Inference* and *Forecast Revision*. Observations with the signal equal to 50 are excluded.

The difference in modal behaviors is an important driver of the inference-forecast gap. The gap shrinks by 36% when we exclude observations with at least one answer classified as exact

[29] In general, extrapolation refers to people's tendency to rely heavily on past outcomes to forecast future outcomes.

[30] In Table B2, we relax the classification criteria for the modes and find similar qualitative patterns. Table B3 shows similar patterns in a participant–part–level classification exercise, where a participant is classified into a type for a given part (*Inference* or *Forecast Revision*) if more than half of her answers in that part are classified into the corresponding mode.

representativeness or naive extrapolation. In a more "reasonable" subsample in which all forecast-revision answers fall within $[0, 100]$ and no answers update in the wrong direction, the inference-forecast gap is in fact reversed when the two modes are excluded, suggesting that the gap is largely explained by the presence of these modes. More details are reported in Tables A6 and A7 of the Appendix.

It is worth noting that all three behavioral modes, albeit capturing different answers, share one common feature: each solely relies on one salient cue in the information environment. Specifically, answers in non-updates, exact representativeness, and naive extrapolation are based entirely on the prior, the expected outcome conditional on the representative state, and the realized signal, respectively. Therefore, instead of properly aggregating all the relevant information, participants simply focus on a few cues—a defining feature of simplifying heuristics (Kahneman and Frederick, 2002; Shah and Oppenheimer, 2008; Gabaix, 2014).

5 Mechanisms

The use of simplifying heuristics *per se* is not surprising given the complexity of the belief-updating tasks. The more surprising observation is that, even when the information environment remains the same, participants end up using different heuristics for solving inference and forecast-revision problems. In this section, we propose and test two complementary mechanisms to explain this divergence in behavior.

5.1 Similarity between cues and elicited variables

5.1.1 Hypothesis development

Building on the literature on salience and memory retrieval (Gennaioli and Shleifer, 2010; Kahana, 2012; Bordalo et al., 2023), we hypothesize that the choice of simplifying heuristics is affected by the similarity between salient cues in the information environment and the variable elicited by the belief-updating question. When the similarity increases, participants are more likely to use that salient cue as an anchor when they form their posterior beliefs.[31]

Table 11 summarizes how similarity can explain the different heuristics observed in *Inference* and *Forecast Revision*. In *Forecast Revision*, the question asks participants to make forecasts about

[31]The memory literature suggests that similarity is a key force in memory recall. In particular, experiences that share common features with the present cue are more "available" to be recalled and therefore have a greater influence on decisions (Kahana, 2012; Bordalo et al., 2023; Jiang et al., 2023). In our setting, cues are that similar to the question could be more likely to enter participants' working memory and therefore affect their beliefs as a salient cue (Afrouzi et al., 2023). This is also related to the compatibility effect (Slovic et al., 1990) which argues that the weight of a cue is enhanced by its compatibility with the response mode.

the stock price growth in the next period conditional on the realized signal. The variable elicited, which is the expected price growth conditioned on the realized signal, shares similarities with the expected price growth conditional on the representative state, as both are values of the outcome variable and are expectations conditioned on the realized signal (in some way). This similarity may lead participants to solely anchor on the expected price growth conditional on the representative state when making forecasts, resulting in exact representativeness. In contrast, *Inference* tasks ask about the conditional probability distribution over the states, which appears less similar to the expected price growth (conditional on the representative state). This may account for the lower frequency of exact representativeness observed in *Inference*.

Table 11: Similarity between belief-updating questions and cues in *Baseline*

Cue	*Inference* $\Pr(\text{state}\vert\text{realized price})$	*Forecast Revision* $\mathbb{E}(\text{price}\vert\text{realized price})$	Behavior
$\mathbb{E}(\text{price}\vert\text{representative state})$	Not similar	Similar	Exact representativeness
Realized price	Not similar	Similar	Naive extrapolation
$\mathbb{E}(\text{price})$		Similar	Non-update
$\Pr(\text{state})$	Similar		Non-update

Similar reasoning can account for the patterns of naive extrapolation observed in the data. In *Forecast Revision*, the realized signal and the elicited variable both pertain to the firm's stock price growth. When participants use the realized signal as an anchor, their forecasts will naively extrapolate from the past. Conversely, the realized signal is less similar to the *Inference* variable, which refers to the conditional probability distribution over the states, and therefore, we rarely observe naive extrapolation in *Inference* problems. Moreover, similarity can also explain the prevalence of non-updates in both types of updating problems. Prior beliefs over states and prior outcome expectations share similarities with their posterior counterparts. When participants anchor on the prior, non-updates may occur.

The mechanism based on similarity also implies that *changes* in the similarity between salient cues and elicited variables should lead to shifts in the prevalence of different simplifying heuristics. To test this prediction, we design two treatments in which we vary the framing of variables and questions to manipulate the similarity between cues and elicited variables.

5.1.2 Evidence from the *More Similar* treatment

In the first similarity treatment, called *More Similar*, we reframe the information environment and the belief-updating questions to *increase* the similarity between the elicited variable in the inference question and the cues. Here, we reframe the signal and the outcome variable as the firm's *profit* in the current month and in the next month, respectively. The state variable is framed as the firm's *profitability*, defined as the long-run average of the firm's monthly profit, taking on values of either 0 or 100. Both the prior distributions of the state and the signal distributions conditional on the state are the same as the corresponding distributions in *Baseline*. The inference question in this treatment asks about the firm's expected *profitability* after the realization of the current month's profit. Similar to *Baseline*, the forecast-revision question asks about the firm's expected profit in the next month conditional on the same signal.

Table 12 summarizes the similarity properties between cues and elicited variables in *More Similar*. Compared to *Baseline*, this treatment increases the similarity between $\mathbb{E}(\text{profitability}|\text{realized profit})$, the variable elicited in the inference question, and two cues: $\mathbb{E}(\text{profit}|\text{representative state})$ and the realized profit. The increase in similarity comes from the fact that the inference variable and the two cues are now all profit-related measures that are conditioned in some way on the realized signal. If participants perceive "profit" and "profitability" as similar concepts, they are likely to use these two cues as anchors for their inference answers, leading to a higher occurrence of exact representativeness and naive extrapolation in inference questions compared to in *Baseline*.

Table 12: Similarity between belief-updating questions and cues in *More Similar*

Cue	*Inference* $\mathbb{E}(\text{profitability}\|\text{realized profit})$	*Forecast Revision* $\mathbb{E}(\text{profit}\|\text{realized profit})$	Behavior
$\mathbb{E}(\text{profit}\|\text{representative state})$	Similar	Similar	Exact representativeness
Realized profit	Similar	Similar	Naive extrapolation
$\mathbb{E}(\text{profit})$		Similar	Non-update
$\mathbb{E}(\text{profitability})$	Similar		Non-update

Table 13 shows that, in *More Similar*, exact representativeness and naive extrapolation become modal behaviors in inference tasks. This is in stark contrast to *Baseline* where these two behaviors are almost non-existent in inference tasks. This treatment also generates a qualitatively different aggregate updating bias from *Baseline* (see Table 14): the fractions of underreacting and overreacting answers are close in *Inference*, and the average update leans towards overreaction. In the Appendix, Table A10 shows in a regression that the inference-forecast gap also becomes smaller but remains marginally significant ($p = 0.079$). This suggests that while reframing the

state variable as a monetary performance measure and asking about its expectation can enhance responsiveness to signals in inference problems, these framing changes do not fully account for the entire inference-forecast gap.

Table 13: Modes of behavior in *More Similar*

Mode	*Inference*	*Forecast Revision*
Non-update	36.3%	28.9%
Exact representativeness	15.3%	18.9%
Naive extrapolation	18.3%	26.9%
No inference-forecast Gap (excluding the other modes)	4.0%	
Unclassified	29.2%	25.5%
Observations	655	655

Notes: The criterion for an answer to be classified into a mode is the same as in Table 10. The percentages are the fractions of answers in each mode. Observations with the signal equal to 50 are excluded.

Table 14: Aggregate patterns in *More Similar*

	Classification			Update
N=86, Obs=655	Underreaction	Near-rational	Overreaction	Mean (s.e.)
Inference	50.7%	6.0%	43.4%	31.0 (4.0)
Forecast Revision	38.5%	4.9%	56.6%	38.0 (3.7)
Rational				23.9 (.5)

Notes: The three columns under "Classification" present the percentages of answers classified as Underreaction, Near-rational, and Overreaction. The last column shows average belief movements in the signal direction from the (objective) priors and their rational benchmark. Observations with the signal equal 50 are excluded. Standard errors are clustered by participant.

5.1.3 Evidence from the *Less Similar* treatment

In a second treatment, called *Less Similar*, we reframe the forecast-revision question so that the elicited variable appears *less* similar to the two salient cues. In this treatment, the state variable, the

signal, and the inference question remain the same as in *Baseline*. We modify the forecast-revision question as follows. After observing the realized stock price growth, participants are asked about *the probability of the firm's revenue going up* next month. The direction of the firm's revenue movement is fully determined by the state—participants are told that a firm's revenue always goes up if the state is Good and down if the state is Bad.

Table 15: Similarity between belief-updating questions and cues in *Less Similar*

Cue	*Inference* $\Pr(\text{state}\|\text{realized price})$	*Forecast Revision* $\Pr(\text{revenue up}\|\text{realized price})$	Behavior
$\mathbb{E}(\text{price}\|\text{representative state})$	Not similar	Not similar	Exact representativeness
$\Pr(\text{revenue up}\|\text{representative state})$	Not salient	Not salient	Exact representativeness
Realized price	Not similar	Not similar	Naive extrapolation
$\mathbb{E}(\text{price})$		Similar	Non-update
$\Pr(\text{state})$	Similar		Non-update

Table 15 examines the similarity properties between cues and elicited variables in *Less Similar*. In this treatment, exact representativeness can arise if participants anchor at one of two cues with values of 100 or 0: $\mathbb{E}(\text{price}|\text{representative state})$, the expected stock price growth conditional on the representative state, and $\Pr(\text{revenue up}|\text{representative state})$, the probability of the revenue going up conditional on the representative state. Compared to in *Baseline*, the first cue $\mathbb{E}(\text{price}|\text{representative state})$ has now become less similar to $\Pr(\text{revenue up}|\text{realized price})$, the variable elicited by the forecast-revision question, as the latter has become a probability distribution over revenue movements. For the second cue $\Pr(\text{revenue up}|\text{representative state})$, although it appears similar to the elicited variable, its values (100% and 0%) are not explicitly stated in the description of the DGP and therefore not as salient as the other cues in the information environment.[32] Therefore, if similarity is driving the use of heuristics, we should see exact representativeness become less prevalent. Analogously, the realized signal (stock price growth in the current month) is no longer similar to the elicited variable (probability of the revenue going up), and we should expect naive extrapolation to show up less.

Table 16 shows the distribution of modal answers in *Less Similar*. Consistent with our prediction, exact representativeness and naive extrapolation are much less prevalent in *Forecast Revision* compared with *Baseline*. This change in modal behaviors supports our hypothesis that when a cue becomes less similar to the question, people are less likely to use heuristics that rely on this cue. Another pattern is that the fraction of answers that satisfy the no inference-forecast gap condition

[32] Specifically, participants are told that "Good firms' revenues grow every month. Bad firms' revenues never grow in any month."

increases from 3.6% in *Baseline* to 11.8% in *Less Similar*. One possible explanation for this result is that the design of *Less Similar* makes it easier for some participants to recognize the tight conceptual connection between inference problems and forecast-revision problems. The change in modal behavior also alters the aggregate pattern of the inference-forecast gap.

Table 17 shows that the inference-forecast gap almost completely vanishes in *Less Similar*, and we obtain the familiar underreaction pattern even in the forecast-revision problems.[33]

Table 16: Modes of behavior in *Less Similar*

Mode	*Inference*	*Forecast Revision*
Non-update	31.7%	30.8%
Exact representativeness	9.0%	13.9%
Naive extrapolation	3.9%	3.6%
No inference-forecast Gap (excluding the other modes)	11.8%	
Unclassified	45.2%	41.5%
Observations	467	467

Notes: The criterion for an answer to be classified into a mode is the same as in Table 10. The percentages are the fractions of answers in each mode. Observations with the signal equal to 50 are excluded.

5.2 Timing of elicited variables

5.2.1 Hypothesis development

The second mechanism we explore concerns the timing of when the elicited variable is realized. Note that, in all our treatments so far, the states are determined before the signals are realized while the forecast outcomes will only be realized in the future (i.e., after the signals). We hypothesize that the relative timing between the realization of states, signals, and outcomes may play a role in the higher prevalence of overreaction-inducing heuristics (such as naive extrapolation) in *Forecast Revision* than in *Inference*, thus contributing to the inference-forecast gap. This hypothesis builds

[33]One may notice that in both *Less Similar* and *Deterministic Outcome* in Section 3.2, the outcome in *Forecast Revision* and the signal are of two different variables. However, forecasts underreact in *Less Similar*, but overreact in *Deterministic Outcome*. These different results can also be reconciled by our hypothesis. Unlike in *Less Similar*, the expected outcome conditional on the representative state remains a salient cue in *Deterministic Outcome* and it is still similar to the elicited forecast variable (the expected outcome conditional on the realized signal). As a result, exact representativeness remains prevalent.

Table 17: Aggregate patterns in *Less Similar*

N=60, Obs=467	Classification			Update
	Underreaction	Near-rational	Overreaction	Mean (s.e.)
Inference	64.7%	12.2%	23.1%	14.3 (1.6)
Forecast Revision	62.1%	12.8%	25.1%	13.6 (1.8)
Rational				23.1 (.6)

Notes: The three columns under "Classification" present the percentages of answers classified as Underreaction, Near-rational, and Overreaction. The last column shows average belief movements in the signal direction from the (objective) priors and their rational benchmark. Observations with the signal equal 50 are excluded. Standard errors are clustered by participant.

on several earlier papers that document a timing effect in decision-making under uncertainty.[34]

5.2.2 Evidence from the *Timing* treatment

To test this timing-based mechanism, we run an additional treatment called *Timing* in which we manipulate the relative timing between signal realization and outcome realization. In *Timing*, participants follow a randomly chosen firm for two consecutive months, labeled the "first month" and the "second month." We randomize participants into two different conditions, the *Future* condition and the *Past* condition.[35] In *Future*, the relative timing between signal and outcome realizations remains the same as in *Baseline*: in each round, participants observe the firm's stock price growth in the first month, and then report either their updated beliefs about the states or their updated expectations about the firm's stock price growth in the second month.[36] In the *Past* condition, however, this relative timing is reversed: after entering the first month, participants are told that the firm's stock price growth in the first month has been determined but is not shown to them. Then, they enter the second month and observe the firm's stock price growth as a signal. Afterwards, they report updated beliefs about the states and about the firm's stock price growth in the first month. Note that our design ensures that the Bayesian benchmarks (under the same signal)

[34]Rothbart and Snyder (1970) and Heath and Tversky (1991) find that people are more willing to bet on realized events than unrealized ones; Nielsen (2020) finds that people prefer earlier resolution of uncertainty for realized events than for unrealized ones; and more relevant to our setting, Benjamin et al. (2017) find that the gambler's fallacy is more pronounced when people predict future coin flips than when they predict past ones, suggesting different belief-formation processes for past and future outcomes.

[35]The full experimental instructions are included in the online appendix.

[36]To finish this round, participants then go through "the second month" where they are told that the firm's stock price growth in the second month has been determined but not shown to them. This makes sure that participants do not receive any feedback throughout the rounds of the experiment.

are always exactly the same across all the problems in the two conditions.

In Table 18, the results from *Future* replicate the results from *Baseline*: participants overwhelmingly underreact to signals in inference problems and overreact to signals in forecast-revision problems. In *Past*, participants also underreact in inference problems, but the degree of overreaction in "precast"-revision problems is much smaller: 53.9% of responses are classified as Overreaction, smaller than the corresponding fraction (64.0%) in *Future* ($p = 0.06$). Similarly, the average update in *Past* (27.9) is also much smaller than the corresponding amount in *Future* (44.9; $p = 0.01$). In the Appendix, Table A12 confirms, in a regression framework, that the inference-precast gap in *Past* is statistically significantly smaller than the inference-forecast gap in *Future*.

Table 18: Aggregate patterns in *Timing*

Future Condition	Classification			Update
N=61, Obs.=470	Underreaction	Near-rational	Overreaction	Mean (s.e.)
Inference	62.3%	12.8%	24.9%	14.0 (1.3)
Forecast Revision	30.0%	6.0%	64.0%	44.9 (5.1)
Rational				23.3 (.6)
Past Condition	Classification			Update
N=59, Obs.=458	Underreaction	Near-rational	Overreaction	Mean (s.e.)
Inference	65.7%	16.2%	18.1%	13.0 (1.4)
"Precast" Revision	40.4%	5.7%	53.9%	27.9 (4.0)
Rational				22.4 (.6)

Notes: We separately report results from the *Future* condition and the *Past* condition. The three columns under "Classification" present the percentages of answers classified as Underreaction, Near-rational, and Overreaction. The last column shows average belief movements in the signal direction from the (objective) priors and their rational benchmark. Observations with the signal equal to 50 are excluded. Standard errors are clustered by participant.

Table 19 further shows that this reduction in the gap is at least partially driven by a decrease in the prevalence of overreaction-inducing heuristics in precast revisions in the *Past* condition. For example, naive extrapolation is significantly less likely to appear in precast revisions compared with in forecast revisions (5.7% vs. 13.2%, $p = 0.01$); exact representativeness is also less likely to appear in precast revisions, although the difference is not statistically significant (21.2% vs. 24.5%, $p = 0.49$).

While a deep dive into the cognitive foundation of the timing effect is beyond the scope of

Table 19: Modes of behavior in *Timing*

	Future Condition		*Past* Condition	
Mode	*Inference*	*Forecast Revision*	*Inference*	*"Precast" Revision*
Non-update	28.3%	14.0%	28.2%	14.4%
Exact representativeness	3.2%	24.5%	1.3%	21.2%
Naive extrapolation	3.4%	13.2%	3.3%	5.7%
No inference-forecast Gap (excluding the other modes)	2.6%		4.1%	
Unclassified	62.6%	48.9%	63.1%	57.0%
Observations	470	470	458	458

Notes: We separately report results from the *Future* condition and the *Past* condition. The criterion for an answer to be classified into a mode is the same as in Table 10. The percentages are the fractions of answers in each mode. Observations with the signal equal to 50 are excluded.

this paper, we do have a conjecture about what could be driving it. Since a realized variable is "set in stone," people may find it more worthwhile to think through all the existing information about it, form a prior belief, and let it "sink in." Once a prior already sinks in, people will be less responsive to new information, which could be driven by confidence (Moreno and Rosokha, 2016) or a preference for commitment (Falk and Zimmermann, 2018). This theory can also explain other timing effects in the literature. For example, confidence in one's belief is often associated with higher willingness to bet (Ellsberg, 1961), which can explain the difference in risk aversion between realized events than unrealized ones (Rothbart and Snyder, 1970; Heath and Tversky, 1991). Additionally, the *ex-ante* eagerness to acquire information about realized uncertainty is directly related to Nielsen's (2020) result that people prefer this kind of uncertainty to be resolved early. Finally, the theory may also speak to the asymmetry between forward- and backward-looking gambler's fallacy in Benjamin et al. (2017) because people who have formed a confident prior about earlier coin flips are less likely to make (biased) inference about them based on new information.

6 Discussion

The behavioral economics literature has documented biased belief updating in various settings. We show, experimentally, that the type of bias uncovered in a specific setting differs between inference and forecast-revision problems. Through a series of experiments, we find that people do

not base their revised forecasts of future outcomes on their own inferences about the underlying states. Instead, many individuals use distinct heuristics to solve the two problems, leading to more overreaction in forecast revision than in inference.

In this section, we discuss primary considerations of external validity of our experimental results, provide suggestive evidence on the use of heuristics in the field, and highlight productive paths for future research.

6.1 External validity

Our experiment documents a disconnect between inference and forecast revision, even in settings where the relationship between states and forecast outcomes is simple and transparent. This disconnect is likely to be even more pronounced in field settings, where the relationship between fundamentals and outcomes is often less clear and complicated by various frictions. For example, in corporate finance, revenue forecasts may depend on multiple factors, such as product popularity, input costs, and competition. If a manager is uncertain about how all these factors jointly determine revenue, then it is difficult to form revenue forecasts based on beliefs about these factors. In the setting of macroeconomic forecasts, the "fundamental state of the economy" is an abstract construct difficult to quantify, let alone infer. These features of field settings make it more likely for forecasts to be disconnected from inference about fundamental states. Indeed, popular corporate finance textbooks such as Welch (2011) propose revenue forecast methods that are disconnected from input forecasts, and recent evidence shows that managers indeed use these methods (Giustinelli and Rossi, 2023).

Our experiments also identify specific heuristics that people use in inference and forecast-revision problems. Naive extrapolation and exact representativeness are primarily used in forecast revision, contributing to overreaction in these tasks. Non-updates, a force of underreaction, are prevalent in both inference and forecast revision. Of course, further research is needed to verify the external validity of these heuristics and discover other decision rules that emerge in specific settings (we provide suggestive evidence on this in the next subsection). However, we have several reasons to believe that the heuristics we identify in our experiment will be important in many field settings. The cues that these heuristics rely on, namely prior beliefs, past realizations and conditional expectations, are often available in the field. Additionally, the factors driving the prevalence of these heuristics, such as similarity and timing, are also present in field settings. Finally, many field settings are more complex than the lab, necessitating the use of simplifying heuristics. Below we further elaborate on the last point, discussing the specific kinds of complexity that may be conducive to the three identified heuristics in field settings.

First, recent research has found that people tend to underreact to complex information (Liang,

2022; Enke and Graeber, 2023). In field settings where new information is hard to interpret, the prior belief is a natural anchor to fall back on. Consequently, non-updates are likely to be prevalent in such settings with complex information.

Second, when the relationship between fundamental states and outcomes is unclear, it is easier to mentally represent outcomes as a univariate time series. This representation naturally leads to the rule of thumb of extrapolating past realizations of the outcome into the future. Giustinelli and Rossi (2023) document this heuristic in managerial forecasts. Kohlhas and Walther (2021) find that when forecasting of a variable, professional forecasters extrapolate from its past realizations, even though they underreact to other information. For inference problems, this heuristic may not be viable because oftentimes the fundamental states of an economy or a company do not have observable past realizations. In addition, as long as the strength of the new signal is not too strong, naively using past realizations to forecast future outcomes often leads to overreaction (Afrouzi et al., 2023).

Third, a large literature has documented the representativeness heuristic both in the lab and in the field, which reflects people's tendency to overweight the state that becomes more likely given the new information (Kahneman and Tversky, 1972; Bordalo et al., 2018, 2019). The heuristic of exact representativeness we identify is an extreme manifestation of this tendency where people focus *exclusively* on the representative state. We suspect that the exact representativeness heuristic is likely to arise in complex field settings because research has shown that integrating consequences from more than one states is difficult (Esponda and Vespa, 2019; Martínez-Marquina et al., 2019). And when forecast revision is more complex than inference, people solving the former task will be more likely to form beliefs based on one state.

6.2 Suggestive evidence from the field

To provide suggestive evidence on the relevance of belief-updating heuristics in the field, we analyze individuals' survey forecasts of two real economic variables: GDP growth rate and stock market returns.

GDP growth rate forecasts of professional forecasters. We first analyze the *Survey of Professional Forecasters* (SPF), which is a quarterly survey of 20-100 professional forecasters conducted by the Federal Reserve Bank of Philadelphia. Following Kohlhas and Walther (2021), we focus on forecasts of quarterly real GDP, which date back to 1968:Q4. Because we do not observe forecasters' mental models about how GDP growth depends on the underlying states of the economy, we cannot identify the heuristic of exact representativeness. Nevertheless, because we observe prior forecasts and past realizations of GDP growth, we can measure the prevalence of non-updates and naive extrapolation. Let y_t denote the year-over-year growth rate of real GDP in quarter t and

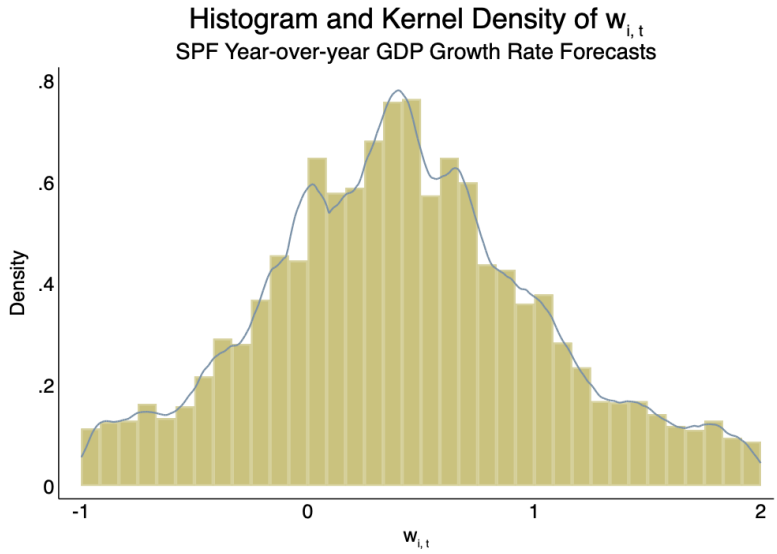

Figure 5: Histogram and kernel density for the weight on the most recent realization (relative to the prior forecast) in the *Survey of Professional Forecasters* (SPF) annual GDP growth rate forecasts, from 1968:Q4 to 2022:Q4. The width of both the bars and the kernel is $\frac{1}{12}$.

let $f_{i,t}y_{t+k}$ denote forecaster i's forecast of y_{t+k} in quarter t. For each one-quarter-ahead forecast $f_{i,t}y_{t+1}$, define

$$w_{i,t} := \frac{f_{i,t}y_{t+1} - f_{i,t-1}y_{t+1}}{y_t - f_{i,t-1}y_{t+1}}, \tag{9}$$

which measures how close the forecast is to the most recent realized GDP growth rate y_t relative to the prior forecast $f_{i,t-1}y_{t+1}$. This measure is equal to 1 if a forecast naively extrapolates from the most recent realization, and it equals 0 if the forecast sticks to the most recent prior.

Figure 5 plots the histogram and kernel density of $w_{i,t}$ in the interval of $[-1, 2]$. The density has a clear spike around 0, suggesting that a significant fraction of GDP growth rate forecasts do not react to recent news – falling into *non-updates*. There is also an excess mass around 1, implying *naive extrapolation*, although its magnitude is much smaller.

Stock market return forecasts of investors. Next, we analyze the UBS/Gallup survey for their *Index of Investor Optimism* (IIO). The IIO is a monthly cross-sectional survey of 1000 investors that ranges from 1998 to 2007. The survey asks respondents to forecast stock market returns in the next 12 months. Again, by comparing this forecast ($f_{i,t}r_{t,t+12}$) to the realized S&P 500 return of the most recent year ($r_{t-12,t}$) and the prior forecast before that, we can identify the heuristics of non-updates and naive extrapolation. Because the IIO is not a panel survey, we do not directly observe a respondent's prior forecast; therefore, we use the annualized S&P 500 return from month

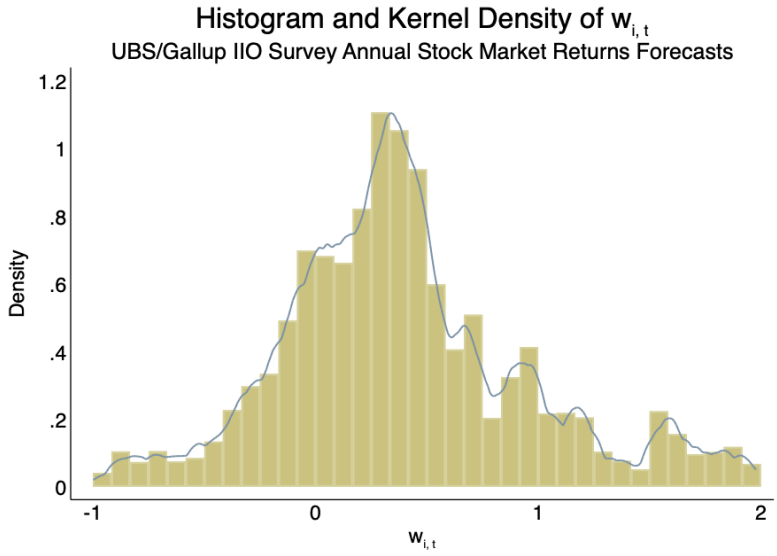

Figure 6: Histogram and kernel density for the weight on the most recent realization (relative to the prior forecast) in annual stock market returns forecasts in the UPS/Gallup *Index of Investor Optimism* (IIO) survey from 1998 to 2007. The width of both the bars and the kernel is $\frac{1}{12}$.

$t-60$ to $t-12$ ($r_{t-60,t-12}$) as a proxy for the prior forecast. Formally, define

$$w_{i,t} := \frac{f_{i,t} r_{t,t+12} - r_{t-60,t-12}}{r_{t-12,t} - r_{t-60,t-12}}, \tag{10}$$

which measures how close the forecast is to the most recent realized market return relative to the proxied prior forecast. Similar to the analysis of the GDP growth rate forecasts, non-updates and naive extrapolation are identified by this measure being close to 0 and 1, respectively.

Figure 6 plots the histogram and kernel density of $w_{i,t}$ in the interval of $[-1, 2]$. The distribution has significant excess masses around both 0 and 1, suggesting that both *non-updates* and *naive extrapolation* are important drivers of stock market return expectations.

6.3 Broader implications

Our study has implications for both theoretical and empirical research on belief-updating biases. One key message of this paper is that individuals can update their beliefs about different variables differently even in the same information environment. This implies a need for both models that allow for internally inconsistent posterior beliefs about different variables and surveys that elicit expectations for multiple variables. Our findings provide a potential resolution to an apparent disconnect between two themes in the literature: the underinference observed in "bookbag-and-

poker-chip" experiments (Benjamin, 2019) and the overinference assumed in models of diagnostic expectations (Bordalo et al., 2018, 2019). Specifically, people may underinfer from new information when they are asked about underlying states, but behave as if they are overinferring when making forecasts as predicted by models of diagnostic expectations.

Our results also demonstrate the prevalence and heterogeneity of belief-updating heuristics, highlighting their importance for aggregate beliefs. Future research could elicit belief-updating heuristics in field settings, potentially through open-ended questions, and incorporate the heterogeneity of these heuristics into economic models.

References

H. Afrouzi, S. Y. Kwon, A. Landier, Y. Ma, and D. Thesmar. Overreaction in Expectations: Evidence and Theory*. *The Quarterly Journal of Economics*, 03 2023. ISSN 0033-5533. doi: 10.1093/qje/qjad009. URL https://doi.org/10.1093/qje/qjad009. qjad009.

P. Andre, C. Pizzinelli, C. Roth, and J. Wohlfart. Subjective models of the macroeconomy: Evidence from experts and representative samples. *Working paper*, 2021.

N. Augenblick, E. Lazarus, and M. Thaler. Overinference from weak signals and underinference from strong signals. *Working paper*, 2023.

C. Ba, J. A. Bohren, and A. Imas. Over-and underreaction to information. *Available at SSRN*, 2022.

N. Barberis, A. Shleifer, and R. Vishny. A model of investor sentiment. *Journal of Financial Economics*, 49(3):307–343, 1998.

N. Barberis, R. Greenwood, L. Jin, and A. Shleifer. X-capm: An extrapolative capital asset pricing model. *Journal of Financial Economics*, 115(1):1–24, 2015.

N. Barberis, R. Greenwood, L. Jin, and A. Shleifer. Extrapolation and bubbles. *Journal of Financial Economics*, 129(2):203–227, 2018.

J. M. Barrero. The micro and macro of managerial beliefs. *Journal of Financial Economics*, 143 (2):640–667, 2022.

D. J. Benjamin. Errors in probabilistic reasoning and judgment biases. *Handbook of Behavioral Economics: Applications and Foundations 1*, 2:69–186, 2019.

D. J. Benjamin, D. A. Moore, and M. Rabin. Biased beliefs about random samples: Evidence from two integrated experiments. Technical report, National Bureau of Economic Research, 2017.

J. B. Berk and R. C. Green. Mutual fund flows and performance in rational markets. *Journal of Political Economy*, 112(6):1269–1295, 2004.

J. R. Bland and Y. Rosokha. Learning under uncertainty with multiple priors: experimental investigation. *Journal of Risk and Uncertainty*, 62(2):157–176, 2021.

P. Bordalo, N. Gennaioli, and A. Shleifer. Diagnostic expectations and credit cycles. *Journal of Finance*, 73(1):199–227, 2018.

P. Bordalo, N. Gennaioli, R. L. Porta, and A. Shleifer. Diagnostic expectations and stock returns. *Journal of Finance*, 74(6):2839–2874, 2019.

P. Bordalo, N. Gennaioli, Y. Ma, and A. Shleifer. Overreaction in macroeconomic expectations. *American Economic Review*, 110(9):2748–82, 2020.

P. Bordalo, N. Gennaioli, A. Shleifer, and S. J. Terry. Real credit cycles. *Working paper*, 2021.

P. Bordalo, J. J. Conlon, N. Gennaioli, S. Y. Kwon, and A. Shleifer. Memory and probability. *The Quarterly Journal of Economics*, 138(1):265–311, 2023.

K. Burdett and T. Vishwanath. Declining reservation wages and learning. *Review of Economic Studies*, 55(4):655–665, 1988.

C. F. Camerer. Do biases in probability judgment matter in markets? experimental evidence. *The American Economic Review*, 77(5):981–997, 1987.

A. Caplin and M. Dean. Revealed preference, rational inattention, and costly information acquisition. *American Economic Review*, 105(7):2183–2203, 2015.

D. L. Chen, M. Schonger, and C. Wickens. oTree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97, 2016.

O. Coibion and Y. Gorodnichenko. Information rigidity and the expectations formation process: A simple framework and new facts. *American Economic Review*, 105(8):2644–78, 2015.

J. J. Conlon, L. Pilossoph, M. Wiswall, and B. Zafar. Labor market search with imperfect information and learning. *Working paper*, 2018.

A. Coutts. Good news and bad news are still news: Experimental evidence on belief updating. *Experimental Economics*, 22(2):369–395, 2019.

D. Ellsberg. Risk, ambiguity, and the savage axioms. *The quarterly journal of economics*, 75(4): 643–669, 1961.

B. Enke. What you see is all there is. *Quarterly Journal of Economics*, 135(3):1363–1398, 2020.

B. Enke and T. Graeber. Cognitive uncertainty. *Working paper*, 2023.

B. Enke and F. Zimmermann. Correlation neglect in belief formation. *Review of Economic Studies*, 86(1):313–332, 2019.

B. Enke, F. Schwerter, and F. Zimmermann. Associative memory and belief formation. *Working paper*, 2021.

L. G. Epstein and Y. Halevy. Hard-to-interpret signals. *Working paper*, 2021.

I. Esponda and E. Vespa. Contingent thinking and the sure-thing principle: Revisiting classic anomalies in the laboratory. *Working paper*, 2019.

I. Esponda, E. Vespa, and S. Yuksel. Mental models and learning: The case of base-rate neglect. Technical report, 2020.

A. Falk and F. Zimmermann. Information processing and commitment. *The Economic Journal*, 128(613):1983–2002, 2018.

S. Fehrler, B. Renerte, and I. Wolff. Beliefs about others: A striking example of information neglect. *Working paper*, 2020.

P. M. Fernbach, A. Darlow, and S. A. Sloman. Asymmetries in predictive and diagnostic reasoning. *Journal of Experimental Psychology: General*, 140(2):168, 2011.

C. Frydman and G. Nave. Extrapolative beliefs in perceptual and economic decisions: Evidence of a common mechanism. *Management Science*, 63(7):2340–2352, 2017.

X. Gabaix. A sparsity-based model of bounded rationality. *Quarterly Journal of Economics*, 129 (4):1661–1710, 2014.

N. Gennaioli and A. Shleifer. What comes to mind? *Quarterly Journal of Economics*, 125(4): 1399–1433, 2010.

N. Gennaioli, Y. Ma, and A. Shleifer. Expectations and investment. *NBER Macroeconomics Annual*, 30(1):379–431, 2016.

T. Gilovich, R. Vallone, and A. Tversky. The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, 17(3):295–314, 1985.

P. Giustinelli and S. Rossi. The coherence side of rationality: Rules of thumb, narrow bracketing, and managerial incoherence in corporate forecasts. 2023.

E. L. Glaeser and C. G. Nathanson. An extrapolative model of house price dynamics. *Journal of Financial Economics*, 126(1):147–170, 2017.

T. Graeber. Inattentive inference. *Working paper*, 2021.

R. Greenwood and A. Shleifer. Expectations of returns and expected returns. *Review of Financial Studies*, 27(3):714–746, 2014.

S. M. Hartzmark, S. Hirshman, and A. Imas. Ownership, learning, and beliefs. *Working paper*, 2021.

S. He and S. Kucinskas. Expectation formation with correlated variables. *Working paper*, 2020.

C. Heath and A. Tversky. Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of Risk and Uncertainty*, 4(1):5–28, 1991.

J. D. Hey. Expectations formation: Rational or adaptive or . . . ? *Journal of Economic Behavior & Organization*, 25(3):329–349, 1994.

Z. Jiang, H. Liu, C. Peng, and H. Yan. Investor memory and biased beliefs: Evidence from the field. *Working paper*, 2023.

M. J. Kahana. *Foundations of human memory*. OUP USA, 2012.

D. Kahneman and S. Frederick. Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and biases: The psychology of intuitive judgment*, 49:81, 2002.

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

C. Kendall and R. Oprea. On the complexity of forming mental models. *Working paper*, 2022.

A. N. Kohlhas and A. Walther. Asymmetric attention. *Working paper*, 2021.

Y. Liang. Learning from unknown information sources. *Working paper*, 2022.

J. Liao, C. Peng, and N. Zhu. Extrapolative bubbles and trading volume. *Working paper*, 2021.

U. Malmendier and S. Nagel. Depression babies: Do macroeconomic experiences affect risk taking? *Quarterly Journal of Economics*, 126(1):373–416, 2011.

U. Malmendier and S. Nagel. Learning from inflation experiences. *Quarterly Journal of Economics*, 131(1):53–87, 2016.

A. Martínez-Marquina, M. Niederle, and E. Vespa. Failures in contingent reasoning: The role of uncertainty. *American Economic Review*, 109(10):3437–3474, 2019.

P. Maxted. A macro-finance model with sentiment. *Working paper*, 2020.

O. M. Moreno and Y. Rosokha. Learning under compound risk vs. learning under ambiguity-an experiment. *Journal of Risk and Uncertainty*, pages 137–162, 2016.

S. Mullainathan. Thinking through categories. *Working paper*, 2002.

S. Mullainathan, J. Schwartzstein, and A. Shleifer. Coarse thinking and persuasion. *Quarterly Journal of Economics*, 123(2):577–619, 2008.

K. Nielsen. Preferences for the resolution of uncertainty and the timing of information. *Journal of Economic Theory*, 189:105090, 2020.

S. Palan and C. Schitter. Prolific.ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018.

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

M. Rabin and D. Vayanos. The gambler's and hot-hand fallacies: Theory and applications. *Review of Economic Studies*, 77(2):730–778, 2010.

M. Rothbart and M. Snyder. Confidence in the prediction and postdiction of an uncertain outcome. *Canadian Journal of Behavioural Science*, 2(1):38, 1970.

A. K. Shah and D. M. Oppenheimer. Heuristics made easy: an effort-reduction framework. *Psychological Bulletin*, 134(2):207, 2008.

C. A. Sims. Implications of rational inattention. *Journal of monetary Economics*, 50(3):665–690, 2003.

P. Slovic, D. Griffin, and A. Tversky. Compatibility effects in judgment and choice. *Insights in decision making: A tribute to Hillel J. Einhorn*, pages 5–27, 1990.

S. Suetens, C. B. Galbo-Jørgensen, and J.-R. Tyran. Predicting lotto numbers: a natural experiment on the gambler's fallacy and the hot-hand fallacy. *Journal of the European Economic Association*, 14(3):584–607, 2016.

A. Tversky and T. Gilovich. The cold facts about the "hot hand" in basketball. *Chance*, 2(1): 16–21, 1989.

A. Tversky and D. Kahneman. Causal schemas in judgments under uncertainty. *Progress in social psychology*, 1:49–72, 1980.

C. Wang. Under-and over-reaction in yield curve expectations. *Working paper*, 2020.

I. Welch. *Corporate finance*. Ivo Welch, 2011.

M. Woodford. Modeling imprecision in perception, valuation, and choice. *Annual Review of Economics*, 12:579–601, 2020.

For Online Publication

A Robustness of the Inference-Forecast Gap

In this section, we examine the properties of the inference-forecast gap in various subsamples of the data.

A.1 A more "reasonable" subsample

We start by examining the inference-forecast gap in a subsample of the *Baseline* treatment that satisfies two basic rationality criteria. In this subsample, we only keep observations whose forecast-revision answer falls within $[0, 100]$, the range bounded by the expected outcome of the Good state and of the Bad state. Furthermore, we exclude observations in which either the inference update or the forecast-revision update is negative; these behavior indicate that the participants' reactions to signals are in the wrong direction.

Table A1: Aggregate patterns in *Baseline*: subsample with "reasonable" updates

N=279, Obs.=1366	Classification			Update
	Underreaction	Near-rational	Overreaction	Mean (s.e.)
Inference	54.5%	17.9%	27.7%	17.7 (.9)
Forecast Revision	42.4%	9.1%	48.5%	24.3 (1.1)
Rational				23.3 (.3)

Notes: The first three columns present the percentages of answers classified as Underreaction, Near-rational, and Overreaction. The last column shows average belief movements in the signal direction from the (objective) priors and their rational benchmark. Observations with the signal equal to 50, forecast-revision answers that are outside $[0, 100]$, or updates in the wrong direction are excluded. Standard errors are clustered by participant.

Table A1 shows the results in this subsample. Although the average update in *Forecast Revision* is close to rational, there is still more overreaction and less underreaction in *Forecast Revision* than in *Inference*. The gap in updates between these two parts is significant, as is shown in a regression analysis in Column (2) of Table A6.

A.2 Priors and signals

The inference-forecast gap exists in all the eight problems with different DGPs (see Table A2). Notably, the eight problems include DGPs with symmetric and asymmetric priors, indicating that our result persists with and without the potential influence of base-rate neglect.

For the subsample with symmetric (objective) priors, we further examine how the inference-forecast gap depends on the strength of the signal. We measure signal strength by the Bayesian update it induces; the more a Bayesian agent moves her belief in response to the signal, the more diagnostic it is about the underlying state. Table A3 shows the results. Overall, there is a larger inference-forecast gap when the signal is more diagnostic, but the gap emerges even for the weakest signals.

Most participants report correct prior beliefs about the states and about the outcome in *Inference Prior* and *Forecast Prior*, but small errors sometimes occur (see Figure C1). To control for the impact of errors in priors on our result, we repeat the classification exercise for the subsample in which the reported inference prior and forecast prior are both correct. The pattern in this sample, shown in Table A4 and in Column (3) of Table A6, is similar: there is more overreaction and less underreaction in *Forecast Revision* than in *Inference*.

A.3 Order between parts

The gap is also robust to different ordering of the five parts. Table A5 compares the gap across different orders and shows that there is a large and statistically significant gap for all three orders. Comparing the inference answers under orders 12345 and 12534 with the forecast revision answers under order 34125, our results also indicate that the gap persists in a between-participant analysis.

A.4 Participant characteristics

We examine the heterogeneity of the gap across participant characteristics, such as gender, education, investment experience, familiarity with statistics and economics, and performance in the comprehension questions. Table A8 show regression results by interacting variables for these characteristics with a *Forecast Revision* dummy. One notable result is that participants who pass all comprehension checks in one pass exhibit less underreaction in *Inference* and less overreaction in *Forecast Revision*, which leads to an inference-forecast gap that is only half as that of the other participants. In addition, participants who report being familiar with economics or finance also exhibit a smaller gap. These results suggest that better comprehension of the subject matter is associated with a smaller inference-forecast gap.

Table A2: Aggregate patterns in *Baseline* (by problem)

		Classification			Update
		Underreaction	Near-rational	Overreaction	Mean (s.e.)
$\Pr(G) = 50\%$	*Inference*	71.1%	19.8%	9.2%	18.7 (1.2)
$\sigma = 50$	*Forecast Revision*	44.3%	12.1%	43.6%	31.2 (2.4)
(Obs. = 273)	Rational				35.9 (.8)
$\Pr(G) = 50\%$	*Inference*	68.2%	16.9%	14.9%	17.0 (1.2)
$\sigma = 60$	*Forecast Revision*	47.9%	6.5%	45.6%	28.6 (2.8)
(Obs. = 261)	Rational				31.8 (.8)
$\Pr(G) = 50\%$	*Inference*	64.8%	13.5%	21.7%	15.3 (1.1)
$\sigma = 70$	*Forecast Revision*	40.8%	7.1%	52.1%	29 (2.6)
(Obs. = 267)	Rational				27 (.8)
$\Pr(G) = 50\%$	*Inference*	64.7%	12.6%	22.7%	13.9 (1.2)
$\sigma = 80$	*Forecast Revision*	40.9%	4.5%	54.6%	34.1 (3.6)
(Obs. = 269)	Rational				25 (.8)
$\Pr(G) = 50\%$	*Inference*	50.6%	18.4%	31.1%	16.2 (1.1)
$\sigma = 90$	*Forecast Revision*	36.7%	4.1%	59.2%	37.3 (3.3)
(Obs. = 267)	Rational				21.8 (.7)
$\Pr(G) = 50\%$	*Inference*	51.3%	16.1%	32.6%	13.1 (1.2)
$\sigma = 100$	*Forecast Revision*	32.2%	8.2%	59.6%	38.3 (3.5)
(Obs. = 267)	Rational				19.7 (.7)
$\Pr(G) = 80\%$	*Inference*	57.4%	13.3%	29.3%	10.6 (1.3)
$\sigma = 100$	*Forecast Revision*	38.1%	3%	58.9%	34.1 (4.1)
(Obs. = 270)	Rational				12.8 (.6)
$\Pr(G) = 20\%$	*Inference*	58.1%	10.7%	31.1%	10 (1.5)
$\sigma = 100$	*Forecast Revision*	36.7%	5.6%	57.8%	29.2 (3.5)
(Obs. = 270)	Rational				12.2 (.5)

Notes: The first three columns present the percentages of answers classified as Underreaction, Near-rational, and Overreaction. The last column shows average belief movements in the signal direction from the (objective) priors and their rational benchmark. Observations with the signal equal to 50 are excluded. Standard errors are clustered by participant.

Table A3: Aggregate patterns in *Baseline* (by signal strength)

Signal Strength		Classification			Update
		Underreaction	Near-rational	Overreaction	Mean (s.e.)
Weakest	*Inference*	47.7%	23.0%	29.3%	4.5 (.8)
(Obs. = 239)	*Forecast Revision*	48.1%	11.7%	40.2%	10.3 (1.5)
	Rational				6.5 (.2)
Weak	*Inference*	59.4%	13.7%	26.8%	9.6 (.9)
(Obs. = 313)	*Forecast Revision*	44.4%	5.4%	50.2%	19.1 (2.1)
	Rational				15.9 (.2)
Medium	*Inference*	63.9%	10.4%	25.7%	15.1 (1.1)
(Obs. = 280)	*Forecast Revision*	37.5%	5.0%	57.5%	33.4 (2.7)
	Rational				25.1 (.1)
Strong	*Inference*	65.0%	12.3%	22.7%	20.4 (1.4)
(Obs. = 300)	*Forecast Revision*	34.7%	4.0%	61.3%	49.6 (4.1)
	Rational				34.4 (.2)
Strongest	*Inference*	63.8%	25.1%	11.0%	25.8 (1.4)
(Obs. = 362)	*Forecast Revision*	43.4%	11.0%	45.6%	39.2 (3.8)
	Rational				44.6 (.2)

Notes: The first three columns present the percentages of answers classified as Underreaction, Near-rational, and Overreaction. The last column shows average belief movements in the signal direction from the (objective) priors and their rational benchmark. Observations with the signal equal to 50 or asymmetric (objective) priors are excluded. The five categories for signal strength correspond to five intervals of rational updates: $[0, 10)$, $[10, 20)$, $[20, 30)$, $[30, 40)$, and $[40, 50]$. Standard errors are clustered by participant.

Table A4: Aggregate patterns in *Baseline*: subsample with correct priors

N=279, Obs.=1502	Classification			Update
	Underreaction	Near-rational	Overreaction	Mean (s.e.)
Inference	57.7%	17.9%	24.4%	15.7 (.8)
Forecast Revision	43.7%	7.7%	48.6%	27.4 (2.3)
Rational				24.1 (.3)

Notes: The first three columns present the percentages of answers classified as Underreaction, Near-rational, and Overreaction. The last column shows average belief movements in the signal direction from the (objective) priors and their rational benchmark. Observations with the signal equal to 50 or with incorrect priors are excluded. Standard errors are clustered by participant.

Table A5: Aggregate patterns in *Baseline* (by order between parts)

		Classification			Update
		Underreaction	Near-rational	Overreaction	Mean (s.e.)
Order: 12345	*Inference*	55.6%	17.5%	27.0%	15.6 (1.1)
(N = 102)	*Forecast Revision*	37.2%	7.3%	55.5%	35.1 (3)
(Obs. = 779)	Rational				22.8 (.4)
Order: 12534	*Inference*	59.9%	16.0%	24.2%	14.5 (1.1)
(N = 103)	*Forecast Revision*	40.4%	5.2%	54.5%	32.4 (3.6)
(Obs. = 795)	Rational				23.0 (.4)
Order: 34125	*Inference*	69.1%	10.9%	20%	12.4 (1.5)
(N = 74)	*Forecast Revision*	42.1%	6.8%	51.1%	29.9 (4.5)
(Obs. = 570)	Rational				24.4 (.5)

Notes: The first three columns present the percentages of answers classified as Underreaction, Near-rational, and Overreaction. The last column shows average belief movements in the signal direction from the (objective) priors and their rational benchmark. Observations with the signal equal to 50 are excluded. Standard errors are clustered by participant.

Table A6: The inference-forecast gap in *Baseline* under various sample restrictions

	Update		
	Full sample	"Reasonable" updates	Correct priors
	(1)	(2)	(3)
Forecast Revision	18.385***	6.682***	11.751***
	(2.279)	(1.210)	(2.530)
Rational Update	1.035***	0.578***	0.926***
	(0.069)	(0.041)	(0.074)
Problem FE	Yes	Yes	Yes
Subject FE	Yes	Yes	Yes
Observations	4288	2732	3004
R^2	0.314	0.463	0.341

Notes: *, **, and *** indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively. Standard errors are clustered by participant. This table presents results for our *Baseline* treatment. Each observation corresponds either to an inference answer or a forecast-revision answer. We define the dependent variable, *Update*, as the answer minus the (objective) prior if the signal is greater than 50, and the opposite if it is smaller than 50. *Rational Update* is the update prescribed by Bayes' rule (and the Law of Iterated Expectations). Observations with the signal equal to 50 are excluded. In Column (2), based on the full sample, we further drop observations with the forecast-revision answer outside the $[0, 100]$ range and observations with at least one update that is in the opposite direction as the signal. In Column (3), based on the full sample, we further drop observations with an incorrect answer for either *Inference Prior* or *Forecast Prior*.

A.5 Framing

Finally, we show that the gap is robust to changing the framing of the signal and forecast outcome. Specifically, in a subsample of the *Baseline* treatment, we frame the signal as the firm's revenue growth (rather than stock price growth); we find a quantitatively smaller but still significant gap with this alternative framing. Table A9 show these results in regressions.

A.6 Regression analyses

Table A7: The inference-forecast gap in *Baseline* excluding modal behaviors

	Update	
	Full sample & excluding two modes	"Reasonable" updates & excluding two modes
	(1)	(2)
Forecast Revision	11.685***	-2.519**
	(2.969)	(1.115)
Rational Update	0.974***	0.446***
	(0.091)	(0.049)
Problem FE	Yes	Yes
Subject FE	Yes	Yes
Observations	2844	1658
R^2	0.321	0.498

Notes: *, **, and *** indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively. Standard errors are clustered by participant. This table presents results for our *Baseline* treatment excluding observations falling into two types of modal behaviors: exact representativeness and naive extrapolation. Each observation corresponds either to an inference answer or a forecast-revision answer. We define the dependent variable, Update, as the answer minus the (objective) prior if the signal is greater than 50, and the opposite if it is smaller than 50. Rational Update is the update prescribed by Bayes' rule (and the Law of Iterated Expectation). Observations with the signal equal to 50 are excluded. In Column (1), based on the full sample, we exclude observations in which the inference answer or the forecast revision answer is classified into one of two modes: exact representativeness and naive extrapolation. In Column (2), we further drop observations with the forecast revision answer outside the $[0, 100]$ range and observations with at least one update that is in the opposite direction as the signal.

Table A8: Heterogeneity of the inference-forecast gap across demographics

	Update
Forecast Revision	30.942***
	(3.849)
Male × *Forecast Revision*	-5.152
	(4.544)
College × *Forecast Revision*	-2.804
	(4.504)
Investor × *Forecast Revision*	-2.310
	(4.516)
Familiar with Stats × *Forecast Revision*	-6.481
	(5.080)
Familiar with Econ × *Forecast Revision*	-6.117
	(5.436)
High Comprehension × *Forecast Revision*	-9.282**
	(3.908)
Male	0.319
	(1.354)
College	-1.031
	(1.447)
Investor	5.009***
	(1.569)
Familiar with Stats	2.634*
	(1.531)
Familiar with Econ	-1.360
	(1.652)
High Comprehension	4.310***
	(1.514)
Rational Update	1.010***
	(0.068)
Problem FE	Yes
Observations	4288
R^2	0.151

Notes: *, **, and *** indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively. Standard errors are clustered by participant. This table presents results for our *Baseline* treatment. Each observation corresponds either to an inference answer or a forecast-revision answer. We define the dependent variable, *Update*, as the answer minus the (objective) prior if the signal is greater than 50, and the opposite if it is smaller than 50. *Rational Update* is the update prescribed by the Bayes' rule (and the Law of Iterated Expectation). Observations with the signal equal to 50 are excluded. We define *Male* as 1 if the participant indicates their gender as Male; the base group is thus Female or Others. We define *College* as 1 if the participant has a bachelor's or postgraduate degree. We define *Investor* as 1 if the participant indicates that they have investments in stocks or mutual funds. We define *Familiar with Stats* as 1 if the participant indicates that they are familiar with probability theory and statistics. We define *Familiar with Econ* as 1 if the participant indicates that they are familiar with economics or finance. We define *High Comprehension* as 1 if the participant correctly answers all the comprehension questions in one pass.

Table A9: Heterogeneity of the Inference-Extrapolation Gap across alternative framing

	Update
Stock Price × *Forecast Revision*	21.187***
	(3.000)
Revenue × *Forecast Revision*	15.654***
	(3.208)
Revenue	1.845
	(1.396)
Rational Update	1.017***
	(0.068)
Problem FE	Yes
Observations	4288
R^2	0.136

Notes: *, **, and *** indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively. Standard errors are clustered by participant. This table presents results for our *Baseline* treatment. Each observation corresponds either to an inference answer or a forecast-revision answer. We define the dependent variable, *Update*, as the answer minus the (objective) prior if the signal is greater than 50, and the opposite if it is smaller than 50. *Rational Update* is the update prescribed by the Bayes' rule (and the Law of Iterated Expectation). Observations with the signal equal to 50 are excluded. Here, we explore heterogeneity of the effects depending on whether we frame the signal as stock price growth or revenue growth.

Table A10: The inference-forecast gap across different treatments

	Update
Baseline × *Forecast Revision*	18.385***
	(2.201)
Deterministic Outcome × *Forecast Revision*	20.697***
	(3.511)
Nudge × *Forecast Revision*	19.708***
	(3.083)
More Similar × *Forecast Revision*	7.009*
	(3.986)
Less Similar × *Forecast Revision*	-0.665
	(1.641)
Deterministic Outcome	-0.881
	(1.475)
Nudge	-3.508**
	(1.460)
More Similar	15.990***
	(4.046)
Less Similar	0.169
	(1.716)
Rational Update	1.029***
	(0.055)
Problem FE	Yes
Observations	9586
R^2	0.139

Notes: *, **, and *** indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively. Standard errors are clustered by participant. In this table, we pool the data from our *Baseline* treatment, *Deterministic Outcome* treatment, *Nudge* treatment, *More Similar* treatment, and *Less Similar* treatment. Each observation corresponds either to an inference answer or a forecast-revision answer. We define the dependent variable, *Update*, as the answer minus the (objective) prior if the signal is greater than 50, and the opposite if it is smaller than 50. *Rational Update* is the update prescribed by the Bayes' rule (and the Law of Iterated Expectation). Observations with the signal equal to 50 are excluded.

Table A11: The inference-forecast gap in *Binary Signal* treatment

	Update
Forecast Revision	3.632*
	(1.992)
Rational Update	0.532***
	(0.074)
Problem FE	Yes
Subject FE	Yes
Observations	2240
R^2	0.204

Notes: *, **, and *** indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively. Standard errors are clustered by particpnt. This table presents results for the *Binary Signal* treatment. Each observation corresponds either to an inference answer or a forecast-revision answer. We define the dependent variable, *Update*, as the answer minus the (objective) prior if the signal is up, and the opposite if it is down. The updates of forecast-revision answers are normalized by $\Pr(\text{up}|G) - \Pr(\text{up}|B)$ so that they are comparable to the inference updates. *Rational Update* is the update prescribed by the Bayes' rule.

Table A12: The inference-forecast gap across the two conditions in the *Timing* treatment.

	Dependent Variable: Update
Forecast Revision	30.943***
	(5.105)
Past Condition × *Forecast Revision*	-15.985**
	(6.303)
Past Condition	-0.124
	(1.961)
Rational Update	1.148***
	(0.105)
Problem FE	Yes
Observations	1856
R^2	0.207

Notes: *, **, and *** indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively. Standard errors are clustered by participant. This table presents results for our *Timing* treatment. Each observation corresponds either to an inference answer or a forecast-revision answer. We define the dependent variable, *Update*, as the answer minus the (objective) prior if the signal is greater than 50, and the opposite if it is smaller than 50. *Rational Update* is the update prescribed by the Bayes' rule (and the Law of Iterated Expectation). Observations with the signal equal to 50 are excluded. We compare the inference-forecast gap in the *Future* condition to the inference-forecast gap in the *Past* condition: Since the *Future* condition is the omitted group, the coefficient before *Forecast Revision* measures the inference-forecast gap in the *Future* condition, and the coefficient before *Past* Condition × *Forecast Revision* measures the reduction in the inference-forecast gap from the *Future* condition to the *Past* Condition.

B Additional Analyses on Modes of Behavior

In this section, we provide additional analyses of the modes of behavior in *Inference* and *Forecast Revision* in the *Baseline* treatment.

B.1 Problems with asymmetric priors

Table B1 quantifies the prevalence of the modal behaviors in problems with asymmetric priors. The overall pattern is similar to that for problems with symmetric priors: non-updates are prevalent in both *Inference* and *Forecast Revision*, while exact representativeness and naive extrapolation show up almost exclusively in the latter.

Table B1: Modes of behavior in *Baseline*: subsample with asymmetric priors

Mode	Criterion for answer	*Inference*	*Forecast Revision*
Non-update	= prior	30.9%	18.1%
Exact Representativeness	= 100 if $s_0 > 50$, = 0 if $s_0 < 50$	2.8%	15.9%
Naive Extrapolation	$= s_0$	3.3%	9.8%
No Inference-Forecast Gap (excluding the other modes)	inference = forecast revision	2.2%	
Unclassified		62.2%	55.4%
Observations		540	540

Notes: The column titled "Criterion for answer" shows the criterion for an answer to be classified into a given mode. Note that an answer may be classified into more than one mode. The percentages in the last two columns are the fractions of answers in each mode in *Inference* and *Forecast Revision* in the *Baseline* treatment. Observations with the signal equal to 50 are excluded.

In forecast-revision problems with symmetric priors, an alternative interpretation of answers classified as exact representativeness is that participants form expectations solely based on the *ex-post more likely* state. This interpretation is distinguishable from the representativeness interpretation in problems with asymmetric priors. For example, consider a forecast-revision problem in which the prior belief $\Pr(G)$ is 20% and the realized signal s_0 is only slightly above 50. Because the signal is good news, the representative state is G. However, because the signal contradicts the prior and is relatively weak, the ex-post more likely state (judged from the participant's own inference) could still be B. Therefore, this problem allows us to differentiate whether participants,

when revising forecasts, are more likely to focus exclusively on the representative state or the ex-post more likely state.

We focus on a subsample of observations in which the objective prior is asymmetric, the reported inference prior and forecast prior are both correct, the signal direction is opposite to the prior direction, and both the inference answer and its rational benchmark are between the prior and 50. Within this subsample, five forecast-revision answers equal the expected outcome of the representative state, whereas none equal the expected outcome of the ex-post more likely state. While the sample size is too small to draw any definitive conclusion, the result nevertheless suggests that participants are more likely to focus on the representative state when they revise forecasts.

B.2 Relaxing criteria for classification

Table B2 shows the prevalence of behavioral modes when we relax the classification criteria to allow for errors within $[-4, 4]$. Compared to the results with strict classification criteria (Table 10), the fraction of answers in each mode increases only slightly, and the overall qualitative pattern remains the same.

B.3 Participant–part–level classification

To study the consistency of behavior within each participant, we conduct a classification exercise at the participant-part level. Specifically, a participant is classified into a type in a part (*Inference* or *Forecast Revision*) if more than half of her answers in that part are classified into the corresponding mode. Table B3 shows the joint distribution of types across the two parts. The numbers of participants classified in the two parts are 73 and 105, and the marginal distribution of types in each part resembles that of the answer-level classification. On the relationship between types in the two parts, many participants are non-updaters in both parts. Meanwhile, participants classified as exact representativeness and naive extrapolation in *Forecast Revision* are mostly unclassified in *Inference*.

B.4 Modes of behavior in other treatments

Table B4 presents results on the modal behaviors in *Deterministic Outcome*. The distribution of modes is similar to *Baseline*. Non-updates are prevalent in both *Inference* and *Forecast Revision*, while exact representativeness and naive extrapolation are only prevalent in the latter.

Table B5 shows that the distribution of modals behaviors in *Binary Signal* are also similar to those in *Baseline*. Non-updates are prevalent in both *Inference* and *Forecast Revision*. In *Forecast Revision*, 17.4% of the answers equal the outcome probability of the representative state, which

Table B2: Modes of behavior in *Baseline* with relaxed criteria for mode classification

Mode	Criterion for answer	*Inference*	*Forecast Revision*
Non-update	$\approx$ prior	32.2%	22.8%
Exact Representativeness	≈ 100 if $s_0 > 50$, ≈ 0 if $s_0 < 50$	5.9%	21.0%
Naive Extrapolation	$\approx s_0$	3.8%	12.1%
No Inference-Forecast Gap (excluding the other modes)	inference $\approx$ forecast revision	3.8%	
Unclassified		54.9%	42.8%
Observations		2144	2144

Notes: The column titled "Criterion for answer" shows the criterion for an answer to be classified into a given mode. The $\approx$ sign means that the criterion allows for errors within $[-4, 4]$. Note that an answer may be classified into more than one mode. The percentages in the last two columns are the fractions of answers in each mode in *Inference* and *Forecast Revision* in the *Baseline* treatment. Observations with the signal equal to 50 are excluded.

constitutes the behavioral mode of exact representativeness. Very few answers are classified as exact representativeness in *Inference*.

Table B6 presents the distribution of modal behaviors in *Nudge*. The fraction of non-updates in *Inference* is 53.2%, a notable increase from the 29.7% in *Baseline*. However, the fraction of non-updates in *Forecast Revision* remains roughly the same as in *Baseline*, as does the fraction of answers classified as exact representativeness and naive extrapolation. In addition, the fraction of answers that satisfy the no inference-forecast gap condition increases to 8.8% from the 3.3% in *Baseline*, suggesting that *Nudge* induces a greater tendency to give internally consistent answers to the two types of updating questions.

Table B3: Joint distribution of *Inference* types and *Forecast Revision* types in *Baseline*

Forecast Revision type \ *Inference* type	Non-update	Exact Representativeness	Naive Extrapolation	No Inference-Forecast Gap	Unclassified	Total
Non-update	22	1	1	0	24	47
Exact Representativeness	2	2	0	0	31	35
Naive Extrapolation	9	0	0	0	12	21
No Inference-Forecast Gap	0	0	0	2	0	2
Unclassified	33	0	1	0	140	174
Total	66	3	2	2	207	279

Notes: This table shows the number of participants that are classified into each type in *Inference* and *Forecast Revision* in the *Baseline* treatment. Note that a participant may be classified into more than one type in a part.

Table B4: Modes of behavior in *Deterministic Outcome*

Mode	Criterion for answer	*Inference*	*Forecast Revision*
Non-update	$=$ prior	35.9%	22.5%
Exact Representativeness	$= 100$ if $s_0 > 50$, $= 0$ if $s_0 < 50$	5.1%	20.8%
Naive Extrapolation	$= s_0$	3.9%	13.3%
No Inference-Forecast Gap (excluding the other modes)	inference = forecast revision	4.6%	
Unclassified		51.5%	42.0%
Observations		777	777

Notes: The percentages in the last two columns are the fractions of answers in each mode in *Inference* and *Forecast Revision* in the *Deterministic Outcome* treatment. Observations with the signal equal to 50 are excluded.

Table B5: Modes of behavior in *Binary Signal*

Part	Mode	Criterion for answer	% of answers
Both	No Inference-Forecast Gap (excluding the other modes)	Equation (7)	2.1%
Inference	Non-update	$\Pr(\theta\|s_0) = Pr(\theta)$	27.1%
	Exact Representativeness	$\Pr(G\|s_0) = 100\%$ if $s_0 =$ up $\Pr(G\|s_0) = 0$ if $s_0 =$ down	3.1%
	Unclassified		67.6%
Forecast Revision	Non-update	$\Pr(s_1\|s_0) = Pr(s_1)$	19.8%
	Exact Representativeness	$\Pr(s_1\|s_0) = Pr(s_1\|G)$ if $s_0 =$ up $\Pr(s_1\|s_0) = Pr(s_1\|B)$ if $s_0 =$ down	17.4%
	Unclassified		60.6%
Observations			1120

Notes: The percentages in the last column are the fractions of answers in each mode for each part in the *Binary Signal* treatment.

Table B6: Modes of behavior in *Nudge*

Mode	Criterion for answer	*Inference*	*Forecast Revision*
Non-update	= prior	53.2%	20.9%
Exact Representativeness	$= 100$ if $s_0 > 50$, $= 0$ if $s_0 < 50$	2.5%	18.0%
Naive Extrapolation	$= s_0$	4.4%	8.8%
No Inference-Forecast Gap (excluding the other modes)	inference = forecast revision	8.8%	
Unclassified		32.8%	45.7%
Observations		750	750

Notes: The percentages in the last two columns are the fractions of answers in each mode in *Inference* and *Forecast Revision* in the *Nudge* treatment. Observations with the signal equal to 50 are excluded.

C Beliefs without realized signal

In this section, we present results from the parts of our experiment in which participants do not see any realized signal: *Inference Prior*, *Forecast Prior*, and *Expectation Formation*. Figure C1 shows the distribution of answers in *Inference Prior* and *Forecast Prior* in the *Baseline* treatment. The majority of answers are correct, with the fraction of correct answers larger under symmetric priors. Participants are more likely to report incorrect priors in *Forecast Prior* than in *Inference Prior*. There are no systematic patterns in the distribution of errors.

Like *Forecast Prior*, the *Expectation Formation* part asks about participants' expectations of the outcome without seeing any realized signal. The unique feature of this part, however, is that the distribution over states in an expectation-formation problem for each participant is set to match the posterior over states reported by this participant in the corresponding inference problem. Figure C2 shows how much expectation-formation answers deviate from the correct answers prescribed by the LoIE in the *Baseline* treatment. The errors are generally small and not large enough to account for much of the inference-forecast gap.

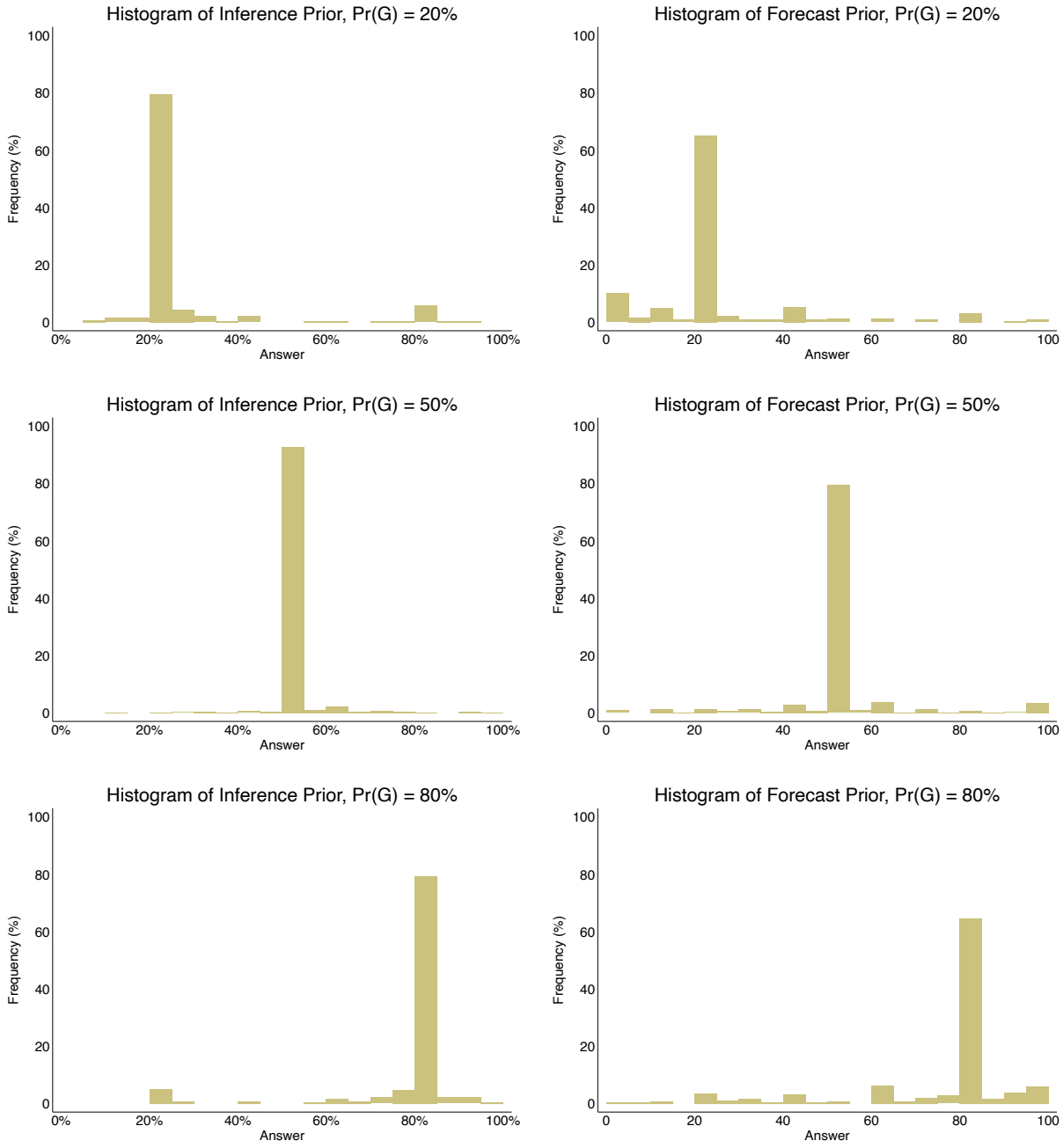

Figure C1: Distributions of answers in *Inference Prior* and *Forecast Prior* in *Baseline*

Notes: We drop a very small fraction of answers in *Forecast Prior* that fall outside [0, 100].

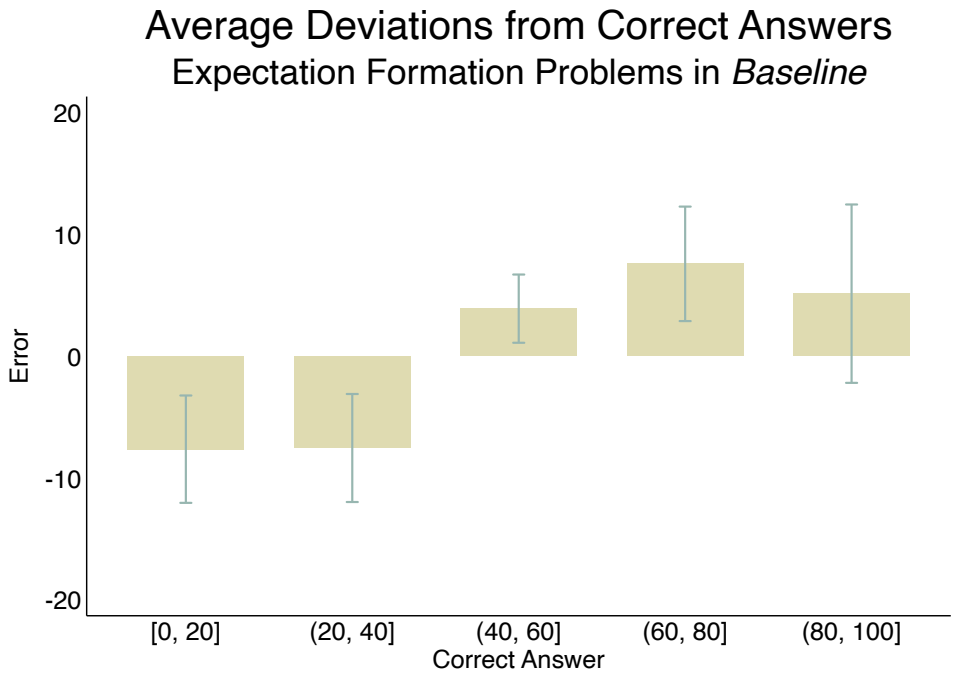

Figure C2: Deviations from LoIE in expectation-formation problems in *Baseline*

Notes: Standard errors are clustered by participant.