

The Inference-Forecast Gap in Belief Updating*

Tony Q. Fan[†]

Yucheng Liang[‡]

Cameron Peng[§]

April 2, 2026

Abstract

Evidence from the laboratory 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. Additional treatments suggest that the choice of heuristics is affected by the similarity between statistics in the information environment and the statistic elicited by the belief-updating problem.

*We thank Peter Andre, Nicholas Barberis, Francesca Bastianello, Daniel Benjamin, B. Douglas Bernheim, Pedro Bordalo, Stefano Cassella, Soo Hong Chew, Marcel Fafchamps, Cary Frydman, Nicola Gennaioli, Matthew Gentzkow, Thomas Graeber, Alex Imas, Spencer Kwon, 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 (while Tony Q. Fan was at Stanford University), 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.

[†]The University of Alabama. Email: qfan5@ua.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 numerous departures from this principle. On the one hand, in the classic “bookbag-and-poker-chip” (or “urn-and-ball”) experiments, participants on average *underreact* to signals when asked to make inferences about the underlying states (see Benjamin, 2019 for a recent review, and Augenblick, Lazarus, and Thaler, 2025; Ba, Bohren, and Imas, 2025 for notable deviations).¹ On the other hand, alternative research designs show that people instead *overreact* to signals when they forecast future outcomes (e.g., Hey, 1994; Frydman and Nave, 2017; AKLM+, 2023). Thus, the direction of belief-updating biases varies across settings and designs. More generally, while economists have long used the notions of underreaction and overreaction to explain major puzzles in macroeconomics and finance (e.g., Barberis, Greenwood, Jin, and Shleifer, 2015; Bordalo, Gennaioli, Shleifer, and Terry, 2026; Maxted, 2024), no consensus has been reached on why people underreact in some environments yet overreact in others (Benjamin, 2019).

In this paper, we examine the determinants of under- and overreaction by analyzing the two commonly studied belief-updating problems side by side. The first is an *inference* problem, where an individual observes informative signals about an underlying state and updates their beliefs about the state. The second is a *forecast-revision* problem, where an individual also observes signals but updates beliefs about future outcomes whose distributions depend on the underlying state. Figure 1 visually illustrates the differences between an inference problem and a forecast-revision problem. For example, in a standard “bookbag-and-poker-chip” experiment, participants are presented with two bookbags containing different proportions of poker chips in different colors. One bookbag is randomly selected, and chips are drawn from it with their colors shown to participants, representing the signals. In an inference problem, participants report their beliefs about which bookbag was chosen. By contrast, in a forecast-revision problem, participants may, for example, predict the color of the next chip to be drawn from the same bookbag.

In both rational and behavioral models, forecast revision is closely tied to inference: Inference about the underlying state typically forms the basis for updating forecasts about future outcomes. However, by running a series of online experiments, we uncover a disconnect between the two problems: on average, participants underreact to signals in the inference problem but overreact in the forecast-revision problem. This gap in belief updating, we show, is largely driven by different simplifying heuristics triggered by the two belief-updating tasks. To explain this gap and the

¹Using standard bookbag-and-poker-chip setups, Augenblick, Lazarus, and Thaler (2025) finds that people overreact when signals are weak and underreact when signals are strong, and Ba, Bohren, and Imas (2025) finds that people overreact when environments are complex, signals are noisy, information is surprising, or priors are concentrated on less salient states.

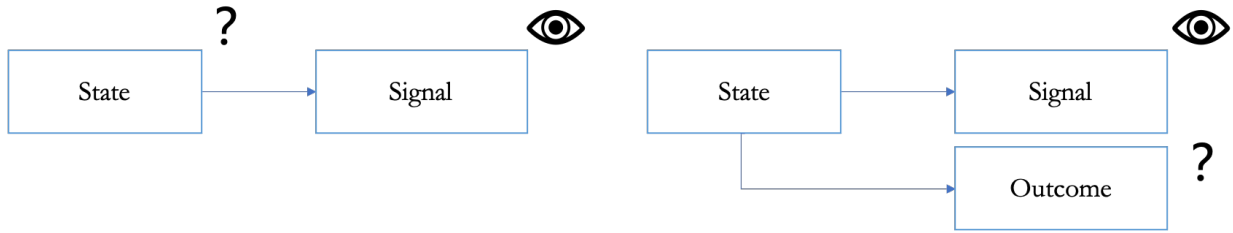


Figure 1: Inference problem (left) and forecast-revision problem (right)

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.

patterns of simplifying heuristics, we propose a conceptual framework based on attribute substitution theory (Kahneman and Frederick, 2002) and run additional treatments to test its predictions. Overall, the results support similarity as a guiding principle for attribute substitution, which leads to different patterns of under- and overreaction.

Our baseline treatment builds on the bookbag-and-poker-chip paradigm in experimental research but frames the relevant variables in economic terms. In each round of the experiment, there is a “firm” with a fixed unknown state which is either “good” or “bad.”² The firm generates signals which are framed as its monthly stock price growth and informative of the state—good firms, on average, have a higher stock price growth than bad firms. Participants are informed of the full data-generating process (DGP), including the prior distribution over the two states and the distributions of signals conditional on each state. Each month, the signal is drawn i.i.d. from a distribution that is discretized from a normal distribution with a mean of 100 if the state is good and 0 if it is bad. This differs from the most common bookbag-and-poker-chip setup where signals follow a binary distribution, a design choice we further discuss below.³

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

²We label the states as “good” and “bad” from the perspective of an investor taking a long position and aiming to maximize their utility or wealth: stocks delivering higher returns are considered “good” for one’s utility, while those delivering lower or negative returns are considered “bad.” More neutrally, a “good” firm could also be described as “high-growth” or “high-return.”

³Section 3.3 presents evidence from a binary-signal treatment and explains why the baseline treatment uses a more continuously distributed signal.

$p\%$ 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' behavior and biases in the two tasks differ qualitatively. 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% of the answers underreact while 54% overreact. Similarly, when belief updates are measured as the belief movement in the signal direction from the prior to the posterior, the average update 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 and 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 several pieces of evidence against this possibility. First, the directions of updating biases are very weakly correlated between *Inference* and *Forecast Revision* at the round or participant levels. Second, signal variance has a significant impact on beliefs in *Inference* but not in *Forecast Revision*. Third, 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-update”: in 30% of the answers, the posterior equals the prior. In *Forecast Revision*, the fraction of

Non-update 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 statistic in the information environment—the prior, the expected outcome conditional on the representative state, and the realized signal—as an anchor in making forecasts. 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 a conceptual framework based on the attribute substitution theory (Kahneman and Frederick, 2002). Our framework posits that, when solving a belief-updating problem that elicits a specific variable (“target statistic”), decision-makers may be uncertain about how to aggregate the available information to compute it. As a result, they may rely on “nontarget statistics” that are (i) easily accessible or computable, and (ii) similar to the target statistic, using these as proxy solutions. For example, in the *Forecast Revision* problem, the target statistic is the expected outcome conditional on the realized signal. Participants who are uncertain about how to calculate this may instead report the expected outcome conditional on the representative state as their answer, as it is both readily available and similar to the target statistic. By contrast, the same nontarget statistic is less similar to the target statistic in the *Inference* problem, which helps explain why Exact Representativeness is less prevalent there. A similar logic applies to Naive Extrapolation: the realized signal is more similar to the target statistic in the *Forecast Revision* problem than in the *Inference* problem, producing a higher incidence of Naive Extrapolation in the former. To formalize these arguments, our framework decomposes any statistic into three components—its statistical measure, essential variable or event, and conditioning event—and defines similarity between two statistics based on their overlap along these components. Using this similarity measure, we then specify a logit model of which nontarget statistic is chosen as a proxy, allowing us to generate quantitative predictions about the distribution of belief-updating responses.

To test the predictions of our framework, we run two additional treatments that vary the similarity between nontarget and target statistics. In one treatment, for instance, we increase the similarity between the two nontarget statistics driving Exact Representativeness and Naive Extrapolation and the target statistic in the *Inference* problem. As predicted, both heuristics become more prevalent, participants react more strongly to signals, and the inference-forecast gap narrows.

Our work is related to an active body of experimental research seeking to understand the conditions of underreaction and overreaction in belief updating (Hartzmark, Hirshman, and Imas, 2021; AKLM+, 2023; Enke and Graeber, 2023; Enke, Schwerter, and Zimmermann, 2024; He and Kučinskas, 2024; Liang, 2025). Recently, Kieren, Müller-Dethard, and Weber (2023), Augenblick, Lazarus, and Thaler (2025), and Ba, Bohren, and Imas (2025) also try to reconcile under- and overreaction in different environments. While we postpone a detailed discussion of the literature to Section 6.1, our focus is to show that the direction of belief-updating biases depends on the type of belief elicited and the nature of the question.⁴ Moreover, by connecting the inference-forecast gap to the use of different simplifying heuristics, we highlight the role of complexity and incorrect mental models in explaining belief-updating biases (Enke and Zimmermann, 2019; Enke, 2020; Andre, Pizzinelli, Roth, and Wohlfart, 2022; Graeber, 2023; Agranov and Reshidi, 2024; Esponda, Vespa, and Yuksel, 2024; Kendall and Oprea, 2024).

The similarity mechanism we propose for the inference-forecast gap builds on classic work on attribute substitution theory (Kahneman and Frederick, 2002) and also draws on insights from recent work on salience and memory retrieval (Kahana, 2012; BCGK+, 2023). In a contemporaneous paper, BCGK+ (2025) develops a model closely related to ours, in which decision-makers attend to salient features of a judgment problem and apply Bayes' rule using only those features to compute probabilities. Our framework is similar in assuming that decision-makers may rely on a reduced set of inputs to form judgments, but the mechanism differs. In our account, decision-makers rely on a single statistic not because they selectively attend to it, but because they are uncertain how to aggregate multiple statistics—such as applying Bayes' rule or the law of iterated expectations. Consequently, rather than incorporating attended features into Bayes' rule, they take the face value of their chosen statistic as a proxy solution. Compared with the model of BCGK+ (2025), our framework can be applied to problems eliciting expectations and other statistics in addition to problems eliciting binary probabilistic beliefs.

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; AKLM+, 2023; He and Kučinskas, 2024), and provides support for overreaction in survey expectations (e.g., Bordalo, Gennaioli, Ma, and Shleifer, 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 feature brings the experimental environment closer to standard

⁴A few belief-updating experiments using the bookbag-and-poker-chip paradigm elicit beliefs about future signals conditional on the current signal (Moreno and Rosokha, 2016; Bland and Rosokha, 2021; Hartzmark, Hirshman, and Imas, 2021; Fehrler, Renerte, and Wolff, 2023; Epstein and Halevy, 2024). None of these experiments compare beliefs about future signals with beliefs about the underlying state (e.g., the bookbag's composition).

models in macroeconomics and finance and affords several advantages for our analysis.⁵ First, the explicit separation between states and outcomes enables us 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 different forms of overreaction, such as representativeness-based overreaction (Kahneman and Tversky, 1972; Bordalo, Gennaioli, and Shleifer, 2018) and naive extrapolation (Barberis, Greenwood, Jin, and Shleifer, 2015, 2018). Third, having a fully specified DGP allows us to attribute biases in posterior beliefs to incorrect statistical reasoning rather than to misperceptions of the DGP.

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.

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, denoted by 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, denoted by 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. $N(0, \sigma^2)$.⁶ 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.

To sum up, the DGP in each round is defined by two elements of information: the prior dis-

⁵In asset-pricing models, investors are typically assumed to learn about firm quality from noisy signals such as stock returns (e.g., Glaeser and Nathanson, 2017). Similarly, in the mutual fund literature, investors learn about manager skill from past fund returns (e.g., Berk and Green, 2004; Rabin and Vayanos, 2010).

⁶In the actual implementation, we discretize the supports of normal distributions to multiples of 10 and truncate at both tails.

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

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

tribution of states and the conditional distributions of signals. These are presented to participants through a one-page display that combines figures and detailed explanations (see Figure 2 for an example and Appendix G for complete screenshots of the experimental interface). Table 2 summarizes the parameter values for the eight DGPs we use, including six with symmetric priors ($\Pr(G) = 50\%$) and two with asymmetric priors. The symmetric-prior DGPs allow us to identify underreaction and overreaction without confounds from base-rate neglect, while the asymmetric-prior DGPs help us examine the robustness of our results. Each DGP appears once in each of the five parts (with modifications in the *Expectation Formation* part, explained later), so answers across parts are directly comparable. Unless stated otherwise, an “observation” refers to a participant’s responses to the five corresponding questions across all five parts.

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

The two main parts of the experiment are *Inference* and *Forecast Revision*. In each round of both parts, participants first observe the firm’s stock price growth in the current month, s_0 . In *Inference*, after observing the realized signal, participants report their updated beliefs about the states $\Pr(\theta|s_0)$. These beliefs are elicited in percentage terms; henceforth, we refer to an inference answer as the reported probability of the Good state (omitting the % sign).⁷ In *Forecast Revision*, participants instead report their updated expectations of the firm’s stock price growth in the next

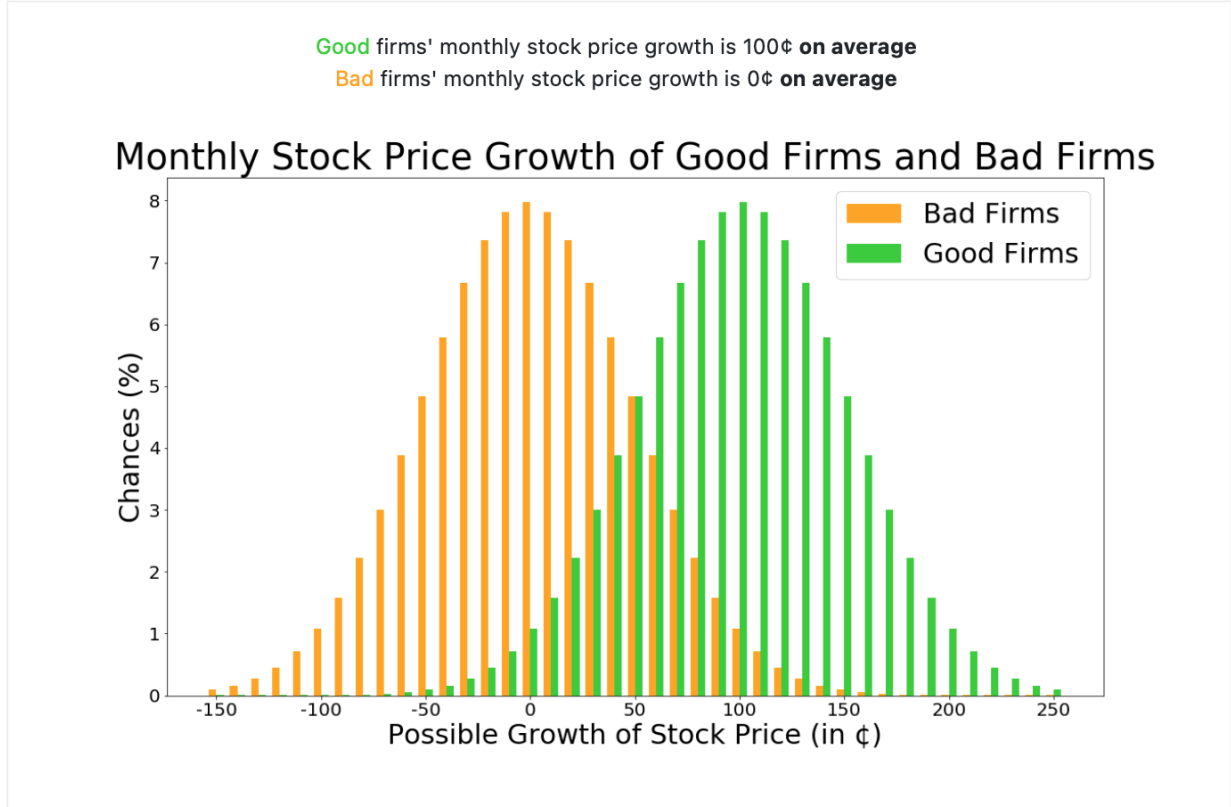
⁷In the experimental interface, there are two blanks: one for the probability of the Good state and one for the Bad state. Entering a number in one blank automatically fills the other with 100 minus that number. Only numbers in the range $[0, 100]$ are allowed.

There is a new pool of 20 firms.

The figure below describes the **stock price growth** of good firms and bad firms in any given month:

The **green** bar on top of each number is the chance (%) that a good firm's stock price grows by that number (in ¢) in any given month.

The **orange** bar on top of each number is the chance (%) that a bad firm's stock price grows by that number (in ¢) in any given month.



The pool of firms has the following composition.



Figure 2: An example of the interface for the DGP

month, $\mathbb{E}(s_1|s_0)$. To enable direct comparison across the two parts, the realized signal is held constant across corresponding rounds for the same participant, though it varies across participants.

In the other three parts, participants do not observe a signal realization before reporting their beliefs. In *Inference Prior*, they directly report prior beliefs about the states $\Pr(\theta)$ based on their knowledge of the DGP. Similarly, in *Forecast Prior*, they directly report prior expectations of the signal $\mathbb{E}(s_1)$. These two parts test whether participants can correctly form prior beliefs. The last part, *Expectation Formation*, mirrors *Forecast Prior* except for the composition of firms in the pool. While the firm composition in *Forecast Prior* is set according to Table 2, in *Expectation Formation* it is based on each participant's posterior beliefs reported in *Inference*. For example, if

a participant reports a posterior belief of $\Pr(G|s_0) = 40\%$ in a given round of *Inference*, then in the corresponding round of *Expectation Formation* the pool of firms will consist of 8 ($= 40\% \times 20$) good firms and 12 bad firms.⁸ This design allows us 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 sufficient attention and discourage click-through behavior, participants are required to remain on each page for at least eight seconds before entering their answers. For each participant, we further randomize (a) the order of the eight DGPs in each part and (b) the order of the five parts. The latter randomization is subject to two constraints: (i) priors have to be elicited before the corresponding posteriors, and (ii) *Expectation Formation* has to follow *Inference*. Hence, only three possible part orders remain: 12345, 12534, and 34125.

After completing the five parts, participants fill out an unincentivized exit survey. At the end of the experiment, they may receive a \$5 bonus payment, with the probability of receiving the bonus determined by their answer in one randomly selected round, according to a quadratic scoring rule.⁹

As a robustness check, we also reframed the signal as *revenue growth* instead of *stock price growth*. As shown in Appendix A.6, the results are qualitatively similar across the two frames, so we pool the data for our main analysis.

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 Expectations (henceforth “LoIE”) implies

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

In our experiment, s_1 and s_0 are independent conditional on the state θ . Hence, $\mathbb{E}(s_1|G, s_0) = \mathbb{E}(s_1|G) = 100$ and $\mathbb{E}(s_1|B, s_0) = \mathbb{E}(s_1|B) = 0$. Substituting into Equation (1) yields

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

⁸If reported beliefs in *Inference* are not multiples of 5%, the numbers of good and bad firms in *Expectation Formation* are rounded to the nearest integer. 14% of the *Inference* answers fall into this case, with half rounded up and half rounded down.

⁹If a participant’s answer in the selected round matches the rational benchmark implied by standard probability theory, they receive the bonus with certainty. Otherwise, the probability of receiving the bonus decreases quadratically with the distance between their answer and the benchmark (see Hartzmark, Hirshman, and Imas, 2021 for a similar incentive structure). Specifically, if the participant’s answer is p and the benchmark is q (both expressed as percentages in the two *Inference* parts), then the bonus probability is $\max\{0, (100 - (p - q)^2)\%$.

We refer to Equation (2) as the “no inference-forecast gap” condition. It captures the theoretical link between posterior beliefs about the state and updated forecasts of the outcome s_1 . When an inference answer and its corresponding forecast-revision answer satisfy this condition, there is no discrepancy between the two belief-updating tasks: Bayesian inference translates into rational forecasts, and any deviation from Bayes’ rule in the inference answer implies 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 posterior of $\Pr(G|s_0) = 40\%$, then the corresponding forecast of s_1 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 generate violations of the no inference-forecast gap condition.¹⁰

When participants solve a forecast-revision problem, one simple and natural procedure that satisfies the no inference-forecast gap condition is what we call the “infer-then-LoIE” procedure. In the first step, participants update their beliefs about the states using the same (possibly non-Bayesian) rule as in the corresponding inference problem. In the second step, they apply the LoIE to these posteriors to form their forecast of s_1 . In later sections, we examine whether participants, in fact, follow this procedure.

2.3 Instructions and comprehension questions

Participants receive extensive instructions, with the tasks and incentive structure explained in detail and in intuitive terms. We place particular emphasis on ensuring that participants fully understand the DGP. First, we emphasize that a firm’s state is constant across months, while signals are i.i.d. conditional on the state. In doing so, we explicitly caution against the incorrect belief that signals are autocorrelated conditional on the state. Second, we use an example DGP to illustrate the discretized normal distributions of the signals, highlighting both the conditional means (0 and 100) and the fact that signals above (below) 50 are good (bad) news about firm quality. Third, we present two explicit formulae: one for computing the prior distribution of states from the pool composition ($\Pr(G) = \frac{\text{number of Good firms}}{20}$) and one for computing the expected signal from the belief about the states ($\mathbb{E}(s) = \Pr(G) \times 100$). However, we do not mention or suggest any specific belief-updating rule.

At the end of the instructions, participants complete a set of comprehension questions testing their understanding of the DGP, the incentive structure, and the two formulae. They could proceed

¹⁰Moreover, because beliefs are equally incentivized across the two types of problems, rational trade-offs between 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.

only after answering all questions correctly.¹¹

2.4 Procedural details

We programmed the experiment using oTree (Chen, Schonger, and Wickens, 2016). For *Baseline*, we recruited 279 participants through Prolific, an online platform for social science research.¹² Signals were framed as monthly revenue growth for 142 participants and as stock price growth for 137 participants. The order of parts also varied: 102 participants experienced the order 12345, 103 the order 12534, and 74 the order 34125. On average, participants spent approximately 30 minutes on the experiment and earned \$7.08, including a \$5 base payment.

2.5 Other treatments

In addition to *Baseline*, we implemented several treatments to examine both the robustness of our results and the underlying mechanisms. An overview of these treatments is provided in Table 3, with details discussed in the respective sections below.

Table 3: Overview of additional treatments

Treatment	Section	Key differences from <i>Baseline</i>
<i>Deterministic Outcome</i>	3.2	Forecast outcome is a different variable (100 if $\theta = G$ and 0 if $\theta = B$)
<i>Binary Signal</i>	3.3	Signals are binary; forecast questions ask about full distributions
<i>Nudge</i>	4.1	Beliefs about states and forecasts are elicited on the same page
<i>More Similar</i>	5.2	State variable (profitability) = mean of signal or forecast outcome (profits); inference questions ask about the expectation of the state
<i>Less Similar</i>	Appendix E	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—

¹¹If mistakes occur, participants are required to re-answer those questions until correct.

¹²See Palan and Schitter (2018) on Prolific as a participant pool. We recruited only U.S. participants who had completed more than 100 tasks on Prolific and maintained an approval rate of at least 99%.

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:

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

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

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

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 2.5.¹³ 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 :

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

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

¹³We 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.

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.

Table 4: Aggregate patterns in *Baseline*

<i>N</i> =279, <i>Obs.</i> =2144	Classification			Update
	Underreaction	Near-rational	Overreaction	Mean (s.e.)
<i>Inference</i>	60.8%	15.2%	24.1%	14.3 (.7)
<i>Forecast Revision</i>	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.

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 A7 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 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. Specifically, the gap’s magnitude shrinks by about two-thirds, suggesting that part of the gap may be driven by differences in how salient the natural bounds are across the two tasks. Specifically, probabilities in *Inference* are naturally bounded within $[0, 100]$, whereas in *Forecast Revision*, although *expectations* should also lie within the same range, *realizations* could exceed it, making the $[0, 100]$ bound less salient and potentially leading some participants to give out-of-bounds answers. We partly address this concern later in the *Deterministic Outcome* treatment by making the bounds more salient in *Forecast Revision*.

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 impact 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, Vallone, and Tversky, 1985; Suetens, Galbo-Jørgensen, and Tyran, 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 in the Good state and 0 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 percentage 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 A11 further confirms, in a regression analysis, that the gap is statistically significant.

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; AKLM+, 2023) in which exaggerated autocorrelation is the key driving force. Moreover, the treatment helps address three 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. Third, by making outcomes deterministic conditional on firm quality, the treatment makes the $[0, 100]$ bound on answers more salient.¹⁴ The fact that the inference-forecast gap remains largely unchanged in this

¹⁴Note that these bounds apply not merely to expectations of future outcomes, but to the *realizations* themselves. Indeed, the fraction of responses outside the $[0, 100]$ range falls significantly.

Table 5: Aggregate patterns in *Deterministic Outcome*

$N=100$, Obs.=777	Classification			Update
	Underreaction	Near-rational	Overreaction	Mean (s.e.)
<i>Inference</i>	64.4%	14.8%	20.8%	13.4 (1.3)
<i>Forecast Revision</i>	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.

treatment suggests that while bound salience—differences in how visible the $[0, 100]$ bounds are across the two tasks—likely contributes to the gap, it cannot be the sole explanation.¹⁵

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. 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%

¹⁵Equalizing bound salience across the two tasks would require making the $[0, 100]$ bound either less salient in *Inference* or more salient in *Forecast Revision*. Both manipulations might reduce the gap, which is consistent with bound salience playing a complementary role in our results. However, making bounds more prominent in *Forecast Revision* would directly suppress the modal behaviors that are the focus of our analysis: in subsequent analyses, we find that *Naive Extrapolation* sometimes involves out-of-range answers, while *Exact Representativeness* implies responses exactly at the boundaries. The persistence of the gap in the *Deterministic Outcome* treatment, where the $[0, 100]$ bound is made more salient, further suggests that bound salience cannot be the sole explanation.

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

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

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

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

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. 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).

Table 7: Aggregate patterns in *Binary Signal*

N=140, Obs.=1120	Classification			Update
	Underreaction	Near-rational	Overreaction	Mean (s.e.)
<i>Inference</i>	61.0%	20.1%	18.9%	11.0 (0.9)
<i>Forecast Revision</i>	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.

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.¹⁶ However, answers in *Forecast Revision* do not exhibit overreaction on average. Overall, the *Binary Signal* treatment shows that the inference-

¹⁶In the Appendix, Table A12 shows in a regression that the gap in updates is significant at the 10% level.

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.

Although the inference-forecast gap persists in the *Binary Signal* treatment, its magnitude is substantially smaller than in *Baseline*. A plausible explanation for this reduction lies in differences in computational and representational complexity across the two treatments. On one hand, as noted earlier, forecast-revision answers are more difficult to compute than inference answers in the *Binary Signal* treatment. This added difficulty may increase cognitive uncertainty (Enke and Graeber, 2023), which could push forecast-revision answers toward underreaction. On the other hand, the space of possible forecast outcomes is much smaller in the *Binary Signal* treatment. Ba, Bohren, and Imas (2025) suggests that when the outcome space is large, people are more prone to overreact by focusing on the outcome most representative of the observed signal. By sharply reducing the number of possible forecast outcomes, the *Binary Signal* treatment likely lowers this complexity, thereby diminishing complexity-induced overreaction in *Forecast Revision*.

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. 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 against the hypothesis that participants intend to follow the infer-then-LoIE procedure but simply make errors in the implementation. 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. Below, we detail these results in succession.

4.1.1 Reducing the computational 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 computational 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.¹⁷ For a participant intending to follow the infer-then-LoIE procedure, this design makes a forecast-revision problem no more computationally 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.¹⁸ For those not intending to follow the infer-then-LoIE procedure, the *Nudge* treatment may implicitly hint at this procedure and encourage participants to adopt it. This could also lead to a reduced inference-forecast gap.

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*.¹⁹

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-LoIE 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 firms.²⁰

¹⁷More 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.

¹⁸Here, we explicitly distinguish between two types of complexity: computational complexity and representational complexity (Ba, Bohren, and Imas, 2025). While the *Nudge* treatment makes the forecast-revision problem less computationally complex, its representational complexity—measured by the number of states and outcomes involved—remains unchanged.

¹⁹In fact, the inference-forecast gap in *Nudge* is even larger than in *Baseline*, according to the regression analysis in Table A11.

²⁰We implement a similar part in *Baseline* as well, and the results are similar (see Section C in the Appendix).

Table 8: Aggregate patterns in *Nudge*

<i>N</i> =100, Obs.=750	Classification			Update
	Underreaction	Near-rational	Overreaction	Mean (s.e.)
<i>Inference</i>	70.9%	10.0%	19.1%	10.1 (1.3)
<i>Forecast Revision</i>	41.3%	6.4%	52.3%	29.8 (3.0)
<i>Expectation Formation</i>	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 belief movement from the (objective) prior in the corresponding forecast-revision problem to the expectation-formation answer in the direction of the forecast-revision signal. The classification of an expectation-formation answer is based on whether its update is smaller, close to, or larger than the rational forecast-revision update. Observations with the signal equal to 50 are excluded. Standard errors are clustered by participant.

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 deviations are small in magnitude across the board. Moreover, in the third row of Table 8, we classify expectation-formation answers and calculate their updates.²¹ Comparing the average update in *Inference*, *Forecast Revision*, and *Expectation Formation*, we find that mistakes in *Expectation Formation* can account for only 23% ($= \frac{14.7-10.1}{29.8-10.1}$) of the inference-forecast gap. Moreover, in Appendix C.2, we illustrate through a quantitative exercise that the degree of overreaction in *Forecast Revision* cannot be matched even by allowing for arbitrarily large mistakes in standalone expectation-formation problems. This is mainly because, in *Inference*, many participants completely ignore the signal and report the 50% prior as their posterior, a pattern we further investigate in Section 4.2. As a result, expectations solely based on these flat posteriors—regardless of how distorted the expectation-formation process is—remain effectively disconnected from the signal and thus cannot account for the overreacting posteriors observed in *Forecast Revision*. Taken together, the results from *Nudge* suggest that the inference-forecast gap does not arise from complexity-induced errors in applying the infer-then-LoIE procedure.

²¹Similar to before, the update of an expectation-formation answer is defined as the belief movement from the (objective) prior in the corresponding forecast-revision problem to the expectation-formation answer in the direction of the forecast-revision signal. The classification of an expectation-formation answer is based on whether its update is smaller, close to, or larger than the rational forecast-revision update.

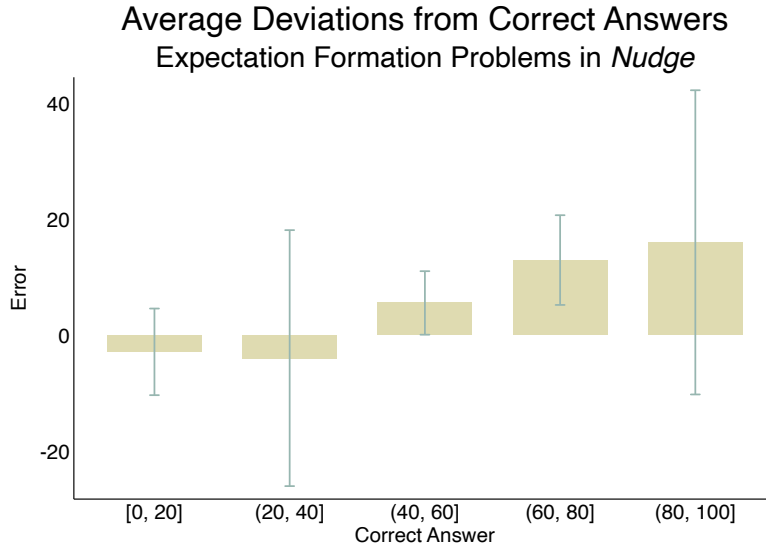


Figure 3: Deviations from LoIE in expectation-formation problems by prior

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.

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*. In the *Baseline* treatment, the correlation between overreaction in inference problems and overreaction in forecast-revision problems is only 0.07 at the round or problem level. At the participant level, the correlation between the fraction of overreactions in *Inference* and the fraction of overreactions in *Forecast Revision* is only 0.12.²² 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 participants follow the infer-then-LoIE procedure. Instead, it suggests that the same individual may use rather different heuristics when answering different types of belief-updating questions. We explore this idea further in Section 4.2.

4.1.3 Reaction to signal variance 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

²²These correlations do increase in the *Nudge* treatment, to 0.27 and 0.25 respectively, which is in line with expectations. However, they are still well below one.

the standard deviation of the signal in updating, separately for *Inference* and *Forecast Revision*. Specifically, we use the following linear specification:²³

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

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 equation (8) separately for Bayesian updates, *Inference* updates, and *Forecast Revision* updates in *Baseline*, 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 Bayesianism implies. Column (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 in *Baseline*?

	Absolute Update: Posterior – Prior		
	Bayesian	<i>Inference</i>	<i>Forecast Revision</i>
	(1)	(2)	(3)
Signal Conditional SD (50 ~ 100)	-0.434 (0.005)	-0.163 (0.024)	-0.016 (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: 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, the results in this section suggest that participants do not appear to be following

²³The 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.

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 answers against realized signals for problems with symmetric objective priors in *Inference* and *Forecast Revision*.²⁴ 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 behavior, which we term “Non-update,” replicates a stylized fact from previous inference experiments (e.g., Coutts, 2019; Graeber, 2023).

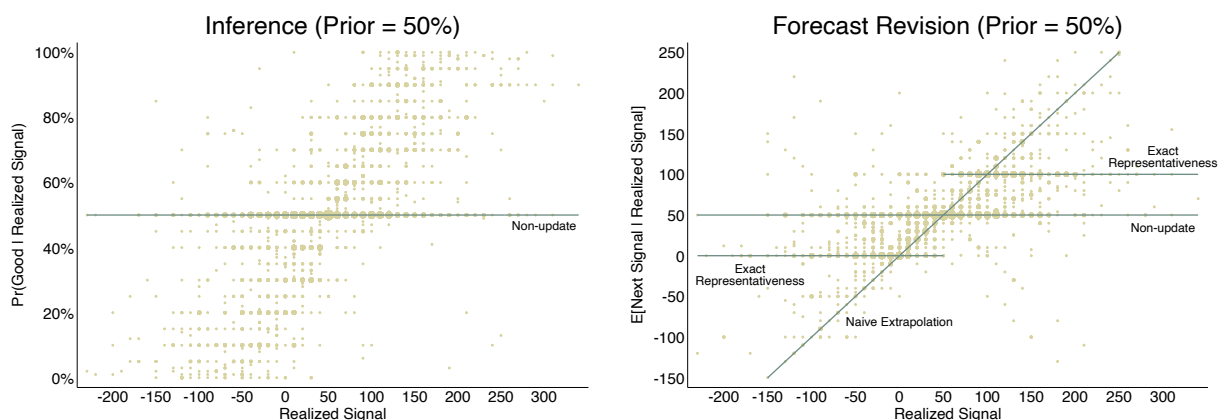


Figure 4: Scatterplots of answers against realized signals: subsample with symmetric priors

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-update also constitutes a mode, shown by a cluster of answers that equal the 50-50 prior. However, two other modes 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 heuris-

²⁴Distributions of answers in problems with asymmetric priors display similar patterns. See Appendix B for details.

tic (Kahneman and Tversky, 1972; Bordalo, Gennaioli, and Shleifer, 2018).²⁵ This behavior is also consistent with a type of belief-updating process induced by coarse thinking (Mullainathan, Schwartzstein, and Shleifer, 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 fits the data better (Mullainathan, 2002).

Second, a smaller yet still significant fraction of forecast-revision answers are anchored at the face value of the realized signal.²⁶ 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, Greenwood, Jin, and Shleifer, 2015, 2018; Liao, Peng, and Zhu, 2022).²⁷ 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-update is 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.²⁸ 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.

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 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

²⁵Note that our notion of Exact Representativeness is different from that in Camerer (1987), who first introduced the term.

²⁶For 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-update and Exact Representativeness are each among the top two most frequent answers for 36 x-axis values.

²⁷In general, extrapolation refers to people’s tendency to rely heavily on past outcomes to forecast future outcomes.

²⁸In Table B3, we relax the classification criteria for the modes and find similar qualitative patterns. Table B4 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.

Table 10: Modes of behavior in *Baseline*

Mode	Criterion for answer	<i>Inference</i>	<i>Forecast Revision</i>
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.

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 A7 and A8 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 falling into Non-update, 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 Mechanism

The use of simplifying heuristics is commonly observed in belief-updating tasks in experimental settings. It is, however, more surprising that participants use different heuristics when solving inference and forecast-revision problems, even though the information environment remains unchanged. To account for these differences, this section develops a simple conceptual framework based on the attribute substitution theory (Kahneman and Frederick, 2002). We then test the new predictions of the framework using additional treatments.

5.1 A similarity-based framework

5.1.1 Setup

Consider a decision-maker (DM) tasked with computing a *target* statistic, denoted by y , based on a description of the environment and the problem. For example, in our inference problem, y corresponds to $\Pr(\theta = G|s_0)$; in our forecast-revision problem, it corresponds to $\mathbb{E}[s_1|s_0]$.²⁹ The description includes 1) other *nontarget* statistics and 2) features pertinent to the calculation of y . In our experiment, examples of nontarget statistics include the numbers of good and bad firms, and the distributions of firm performance conditional on firm quality. Features include both additional information on the properties of nontarget statistics, which we detail later, and structural aspects of the DGP, such as the assumption that firm performance across months is i.i.d. conditional on firm quality.

The DM then applies a procedure to estimate y based on her understanding of the problem and the environment. However, the estimate of the target statistic, denoted by \hat{y} , may be incorrect: the DM may either apply an incorrect procedure or make errors in implementing the correct one. We assume that the DM, who is at least partially aware of her potential inaccuracy (Enke and Graeber, 2023), perceives \hat{y} as a noisy proxy of y , modeled as $\hat{y} = y + \epsilon_{\hat{y}}$, where $\epsilon_{\hat{y}}$ is an error term with variance $\text{Var}(\epsilon_{\hat{y}}) = \frac{1}{\tau_{\hat{y}}}$ and $\tau_{\hat{y}}$ denotes the perceived precision of \hat{y} as an estimate of y . If the DM is fully confident in her computation (i.e., $\tau_{\hat{y}} = \infty$), she will use \hat{y} as her final answer. Otherwise, she may consider expanding the set of candidate answers beyond \hat{y} .

Two types of nontarget statistics naturally serve as candidate answers. First, those readily available in the description, such as s_0 and $\mathbb{E}[s|\theta = G]$. Second, those easily computable from the description, such as $\Pr(\theta = G)$ and $\mathbb{E}[s]$.³⁰ We denote all nontarget statistics by x_1, x_2, \dots, x_n . The DM understands that nontarget statistics are conceptually different from y . Accordingly, as with \hat{y} , she perceives each x_i as a noisy proxy of y : $x_i = y + \epsilon_i$, where $\text{Var}(\epsilon_i) = \frac{1}{\tau_{x_i, y}}$ for $i = 1, 2, \dots, n$.

We further assume that the DM cannot aggregate the $n + 1$ statistics to form a better estimate of y . This assumption is psychologically realistic, as people often do not know how to combine multiple pieces of information to form an accurate estimate (Drugowitsch, Wyart, Devauchelle, and Koechlin, 2016). This assumption is also without loss of generality: if the DM were able to aggregate the statistics, we could simply redefine \hat{y} as the newly aggregated estimate and proceed with the same framework.³¹

²⁹In the two problems eliciting priors, the target statistic y corresponds to the unconditional counterparts, $\Pr(\theta = G)$ and $\mathbb{E}[s_1]$, respectively.

³⁰In principle, the DM may also make mistakes in computing the second type of statistics. However, we abstract away from such mistakes for simplicity in this model.

³¹In particular, this implies that the framework can accommodate combinations of nontarget statistics: an agent who weights multiple statistics could be reinterpreted as one whose \hat{y} already incorporates that combination. Empirically, unclassified answers, which display more continuous patterns, are consistent with noisy computation of the target

5.1.2 The structure of similarity

What determines the perceived precision (τ_i) in the framework described above? Building on the growing literature on the role of similarity in cognition and belief formation (Logan, 2021; BCGK+, 2023; Jiang, Liu, Peng, and Yan, 2025), we assume that a nontarget statistic is perceived as a more precise signal of the target statistic when the two appear more similar. To quantify similarity, note that each statistic—whether target or nontarget—is characterized by three distinct dimensions:

1. Statistical measure m : In our setting, this is either the expectation operator \mathbb{E} or the probability measure Pr . More generally, it could be other measures such as quantiles or variance.
2. Essential variable or event e : The variable or event that the statistical measure operates on.
3. Conditioning event c : The event assumed to have occurred when calculating the statistic.

For instance, consider the two target statistics in the inference and forecast problems: $\mathbb{E}[s_1|s_0]$ and $\text{Pr}(\theta = G|s_0)$. They differ in both their statistical measures and essential variables: the first is the mathematical expectation (\mathbb{E}) of a future outcome (s_1), while the second is the probability (Pr) of firm quality ($\theta = G$).³² However, they share the same conditioning event, namely the realized signal s_0 . More broadly, this three-dimensional characterization—statistical measure, essential variable, and conditioning event—provides a general framework for describing and comparing a wide range of statistical objects across different contexts. For example, the framework can be extended to unconditional statistics by treating their conditioning event as the full set. Likewise, statistics such as s_0 can be reformulated as $\mathbb{E}[s_0|s_0]$.

Consider two statistics x ($x = \{m_x, e_x, c_x\}$) and y ($y = \{m_y, e_y, c_y\}$). We assume that their overall similarity, denoted by $\Delta(x, y)$, aggregates their similarities across the three dimensions. Formally,

$$\Delta(x, y) = \phi(\delta_m(m_x, m_y), \delta_e(e_x, e_y), \delta_c(c_x, c_y)), \quad (9)$$

where ϕ is an aggregation function and δ_k ($k \in \{m, e, c\}$) represents the pairwise similarity along a given dimension. Rather than specifying the exact forms of these functions, we impose the following two general conditions:

statistic or with partial anchoring on a nontarget statistic.

³²One might ask whether the distinction between \mathbb{E} and Pr as separate statistical measures is well-defined, since one can always write $\text{Pr}(\theta = G) = \mathbb{E}[\mathbf{1}(\theta = G)]$. Our framework is defined over participants' natural perceptual representations of statistics, not over all mathematically equivalent formulations. Participants naturally perceive the inference problem as asking for a probability and the forecast-revision problem as asking for an expectation. Accordingly, treating \mathbb{E} and Pr as distinct statistical measures is the psychologically appropriate choice. Note also that this rewrite simultaneously changes both the statistical measure and the essential variable (from $\theta = G$ to $\mathbf{1}(\theta = G)$), and so does not straightforwardly reduce to a relabeling of the measure dimension alone.

Condition 1. (*Monotonicity*) The aggregator function ϕ is increasing in each of its arguments.

Condition 2. (*Maximum-at-Identity*) Each similarity function δ_k reaches its maximum when the two statistics are identical in dimension k .

In addition to these two formal conditions, we sometimes also assume, informally, that δ_e captures perceptions of semantic similarity. For example, two variables are perceived to be similar if they share the same variable name, unit of measurement, or statistical distribution.

The structure for measuring similarity introduced above can be directly applied to our experiment to explain the emergence of the three simplifying heuristics: Exact Representativeness (ER), Naive Extrapolation (NE), and Non-update (NU). Consider a scenario in which the realized signal represents good news, that is, $s_0 > 50$. First, the nontarget statistic associated with ER, denoted by x_{ER} , is $\mathbb{E}[s_1 | \theta = G]$. Comparing it to the target statistic in *Forecast Revision*, $y^{\text{FR}} = \mathbb{E}[s_1 | s_0]$, the two share the same statistical measure \mathbb{E} and the same essential variable s_1 . By contrast, comparing x_{ER} to the target statistic in *Inference*, $y^{\text{I}} = \Pr(\theta = G | s_0)$, they differ in both the statistical measure and the essential variable. Given that x_{ER} is equally similar to the two target statistics in the third dimension, i.e., the conditioning event, Conditions 1 and 2 imply that x_{ER} is more similar to y^{FR} than to y^{I} .

Second, the nontarget statistic associated with NE, s_0 , can be reformulated as $x_{\text{NE}} = \mathbb{E}[s_0 | s_0]$, which shares the same statistical measure with y^{FR} but not with y^{I} . The essential variable of x_{NE} (s_0) is also arguably more similar to the essential variable of y^{FR} (s_1) than to that of y^{I} (θ), since s_0 and s_1 share the same variable name, unit of measurement, and unconditional distribution. Consequently, the nontarget statistic x_{NE} is more similar to y^{FR} than to y^{I} .³³

Third, in the case of NU, each target statistic— $y^{\text{I}} = \Pr(\theta = G | s_0)$ or $y^{\text{FR}} = \mathbb{E}[s_1 | s_0]$ —shares the same statistical measure and essential variable with its unconditional/prior counterpart— $x_{\text{NU}}^{\text{I}} = \Pr(\theta = G)$ or $x_{\text{NU}}^{\text{FR}} = \mathbb{E}[s_1]$, respectively—and is thus highly similar to the latter. Moreover, by the *Maximum-at-Identity* condition, the aggregate similarity measure between the prior and the posterior is identical across the two belief-updating problems. Table 11 summarizes the dimension-by-dimension comparisons.

³³The reformulation of s_0 as $\mathbb{E}[s_0 | s_0]$ is technically necessary to fit s_0 into our 3-dimensional similarity framework. Psychologically, the fact that the realized signal is similar to the target statistic in forecast revision in terms of the essential variable and conditioning event is sufficient for the logic to go through.

Table 11: Similarity between nontarget and target statistics across dimensions

Heuristic	Nontarget statistic	<i>Forecast Revision</i> target: $\mathbb{E}[s_1 s_0]$			<i>Inference</i> target: $\Pr(\theta = G s_0)$		
		Measure	Ess. var.	Cond. event	Measure	Ess. var.	Cond. event
ER	$\mathbb{E}[s_1 \theta = G]$	✓	✓	×	×	×	×
NE	$\mathbb{E}[s_0 s_0]$	✓	≈	✓	×	×	✓
NU	$\mathbb{E}[s_1] / \Pr(\theta = G)$	✓	✓	×	✓	✓	×

Notes: ✓ denotes a match; × denotes a mismatch; ≈ denotes semantic similarity (s_0 and s_1 share the same variable name, unit, and unconditional distribution). For NU, the nontarget statistic is $\mathbb{E}[s_1]$ in *Forecast Revision* and $\Pr(\theta = G)$ in *Inference*.

5.1.3 Similarity and the frequencies of modal answers

Having specified the structure of similarity, we next introduce a logit specification to characterize behavior. We begin with the following condition:

Condition 3. (*Logit*)

- The perceived precision of x_i as a signal for y , denoted by τ_{x_i} , follows a Gumbel distribution with scale parameter β and location parameter $\ln(\Delta(x_i, y))$, where $\Delta(x_i, y)$ is the similarity between x_i and y .
- The perceived precision of \hat{y} as a signal for y , denoted by $\tau_{\hat{y}}$, follows a Gumbel distribution with scale parameter β and location parameter $\ln(\mu_{\hat{y}})$.
- The distributions of $\tau_{\hat{y}}$ and the τ_{x_i} 's are independent.³⁴

Result 1. Under Condition 3, the frequency of answers for y that equal x_i is given by

$$\pi(x_i, y) = \frac{\Delta(x_i, y)}{\mu_{\hat{y}} + \sum_j \Delta(x_j, y)}, \quad (10)$$

³⁴Condition 3 is a standard parametric assumption from discrete choice theory: treating the perceived precision of each candidate answer as a random variable with a Gumbel distribution yields a logit formula for frequencies, chosen here for tractability. The key qualitative prediction that frequency is increasing in similarity holds under any specification in which higher similarity stochastically shifts the precision distribution upward.

and is increasing in $\Delta(x_i, y)$, ceteris paribus.

According to Result 1, the likelihood of choosing a nontarget statistic increases with its perceived similarity to the target. Recall that the discussion of target–nontarget similarity in *Inference* and *Forecast Revision* implies

$$\Delta(x_{\text{NU}}^{\text{I}}, y^{\text{I}}) = \Delta(x_{\text{NU}}^{\text{FR}}, y^{\text{FR}}), \quad \Delta(x_{\text{ER}}, y^{\text{I}}) < \Delta(x_{\text{ER}}, y^{\text{FR}}), \quad \text{and} \quad \Delta(x_{\text{NE}}, y^{\text{I}}) < \Delta(x_{\text{NE}}, y^{\text{FR}}).^{35}$$

Result 1 allows us to make the following prediction about different rates of simplifying heuristics across *Inference* and *Forecast Revision*.

Prediction 1. Assume that

$$\mu_{\hat{y}}^{\text{I}} \geq \mu_{\hat{y}}^{\text{FR}}.$$

Define the proportional difference in target–nontarget similarity between *Inference* and *Forecast Revision* as

$$\text{Diff}_{\text{NE}} = \frac{\Delta(x_{\text{NE}}, y^{\text{FR}}) - \Delta(x_{\text{NE}}, y^{\text{I}})}{\Delta(x_{\text{NE}}, y^{\text{I}})}, \quad \text{Diff}_{\text{ER}} = \frac{\Delta(x_{\text{ER}}, y^{\text{FR}}) - \Delta(x_{\text{ER}}, y^{\text{I}})}{\Delta(x_{\text{ER}}, y^{\text{I}})}.$$

and the tolerance ratios as

$$R_{\text{ER}} = \frac{\mu_{\hat{y}}^{\text{I}} + \Delta(x_{\text{NU}}, y^{\text{I}})}{\Delta(x_{\text{ER}}, y^{\text{I}})}, \quad R_{\text{NE}} = \frac{\mu_{\hat{y}}^{\text{I}} + \Delta(x_{\text{NU}}, y^{\text{I}})}{\Delta(x_{\text{NE}}, y^{\text{I}})}.^{36}$$

If the proportional increases satisfy

$$\frac{1}{1 + R_{\text{ER}}} < \frac{\text{Diff}_{\text{NE}}}{\text{Diff}_{\text{ER}}} < 1 + R_{\text{NE}},$$

then

$$\pi(x_{\text{ER}}, y^{\text{I}}) < \pi(x_{\text{ER}}, y^{\text{FR}}), \quad \pi(x_{\text{NE}}, y^{\text{I}}) < \pi(x_{\text{NE}}, y^{\text{FR}}).$$

In words, Prediction 1 says that both ER and NE should be more prevalent in *Forecast Revision* than in *Inference*. The first condition ($\mu_{\hat{y}}^{\text{I}} \geq \mu_{\hat{y}}^{\text{FR}}$) reflects that the DM is less confident in her direct computation in *Forecast Revision*, leaving more room for heuristics overall. The second condition—captured by the tolerance ratio bounds—requires that the similarity gains are roughly balanced across the two heuristics: since ER and NE compete for the same “market share” of

³⁵Note that $x_{\text{NU}}^{\text{I}} = \Pr(\theta = G)$ and $x_{\text{NU}}^{\text{FR}} = \mathbb{E}[s_1]$ are two different statistics, although they share the same numerical value in our experiment.

³⁶The tolerance ratio R_k captures the competitive pressure that heuristic k faces in *Inference*: its numerator is the combined baseline appeal of two alternatives to k (the computed answer and non-updating), while its denominator is k 's own baseline appeal.

responses, if one gains far more than the other in moving from *Inference* to *Forecast Revision*, it could crowd out the other’s frequency despite both becoming individually more attractive. See Appendix B.6 for a proof.

Prediction 1 is consistent with the experimental evidence documented in Table 10. By the same logic, the model also predicts that if we simultaneously increase (or decrease) the similarity between the two nontarget statistics and the target statistic, Exact Representativeness and Naive Extrapolation will both become more (or less) prevalent, provided the proportional changes are not too different between the two nontarget statistics. This prediction is formalized below.

Prediction 2. For $y \in \{y^I, y^{FR}\}$, suppose $\Delta(x_{ER}, y)$ increases (decreases) to $\Delta'(x_{ER}, y)$ and $\Delta(x_{NE}, y)$ increases (decreases) to $\Delta'(x_{NE}, y)$. Define the proportional changes as

$$\text{Change}_{NE} = \frac{\Delta'(x_{NE}, y) - \Delta(x_{NE}, y)}{\Delta(x_{NE}, y)}, \quad \text{Change}_{ER} = \frac{\Delta'(x_{ER}, y) - \Delta(x_{ER}, y)}{\Delta(x_{ER}, y)}.$$

and the tolerance ratios as

$$R_{ER} = \frac{\mu_{\hat{y}} + \Delta(x_{NU}, y)}{\Delta(x_{ER}, y)}, \quad R_{NE} = \frac{\mu_{\hat{y}} + \Delta(x_{NU}, y)}{\Delta(x_{NE}, y)}.$$

If the proportional changes satisfy

$$\frac{1}{1 + R_{ER}} < \frac{\text{Change}_{NE}}{\text{Change}_{ER}} < 1 + R_{NE}$$

then $\pi(x_{ER}, y)$ and $\pi(x_{NE}, y)$ —the frequencies of Exact Representativeness and Naive Extrapolation—will also increase (decrease) accordingly.

5.2 Additional evidence for the similarity-based framework

To provide additional causal evidence for the similarity-based framework, we directly test Prediction 2 by conducting two new treatments that manipulate the similarity between the target statistic of a belief-updating problem and the overreaction-inducing nontarget statistics—namely those associated with Exact Representativeness and Naive Extrapolation.

In the first new treatment, which we call *More Similar*, we reframe the information environment and belief-updating problems to *increase* the similarity between the target statistic y in *Inference* and the overreaction-inducing nontarget statistics. Specifically, 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 monthly profit and taking values of 0 or 100. Both the prior distribution of the state and the signal distributions

conditional on each state are identical to those in *Baseline*. In this treatment, the inference problem asks participants to report the firm’s expected *profitability* conditional on its realized profit in the current month. Similar to *Baseline*, the forecast-revision problem asks for the firm’s expected profit in the next month given the same signal.

Compared to *Baseline*, this treatment increases the similarity between the target statistic elicited in the inference problem—here $\mathbb{E}[\text{profitability}|\text{realized profit}]$ —and two nontarget statistics: the realized profit and $\mathbb{E}[\text{profit}|\text{representative state}]$. The increase in similarity arises because all these statistics can be represented as conditional expectations of profit-related variables—assuming that “profit” and “profitability” are perceived as similar variables. Following Prediction 2, we expect participants to more frequently use these two nontarget statistics as their answers to the inference problem, leading to a higher incidence of Exact Representativeness and Naive Extrapolation in *Inference* relative to *Baseline*.

Table 12 shows that, in *More Similar*, Exact Representativeness and Naive Extrapolation indeed emerge as modal behaviors in *Inference*. This stands in sharp contrast to *Baseline* where these two behaviors are almost nonexistent in *Inference*. This treatment also generates an average updating bias that is qualitatively different from that in *Baseline* (see Table 13): the fractions of underreacting and overreacting responses are close in *Inference*, and the average update tilts towards overreaction. As shown in Appendix Table A11, the inference-forecast gap becomes smaller in *More Similar*, although it remains marginally significant ($p = 0.079$). Overall, the similarity manipulation substantially increases belief updating in *Inference*, even though it does not fully eliminate the inference-forecast gap.³⁷

In the second new treatment, which we call *Less Similar*, we reframe the forecast-revision problem to *reduce* the similarity between its target statistic and the overreaction-inducing nontarget statistics. In *Less Similar*, both Naive Extrapolation and Exact Representativeness become substantially less prevalent in the forecast-revision problem relative to *Baseline*. Consequently, the inference-forecast gap nearly disappears. Detailed results are reported in Appendix E.

6 Discussion

In this section, we discuss relationships with other accounts of over- and underreaction in the literature, the generality of our theoretical framework, and the external validity of our results.

³⁷Why does a residual inference-forecast gap remain in *More Similar*? There is suggestive evidence that the temporal nature of the elicited statistic may explain it. Specifically, Appendix D shows that participants are less likely to overreact when updating their beliefs about outcomes realized *before* the observed signal than when updating about outcomes to be realized in the future.

Table 12: Modes of behavior in *More Similar*

Mode	<i>Inference</i>	<i>Forecast Revision</i>
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 13: Aggregate patterns in *More Similar*

	Classification			Update
	Underreaction	Near-rational	Overreaction	Mean (s.e.)
N=86, Obs=655				
<i>Inference</i>	50.7%	6.0%	43.4%	31.0 (4.0)
<i>Forecast Revision</i>	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.

6.1 Relationships to other accounts of over- and underreaction

While both under- and overreaction have been extensively documented (e.g., Benjamin, 2019; Bordalo, Gennaioli, Ma, and Shleifer, 2020), only recently has the experimental literature begun to systematically examine their conditions and mediators. Specifically, our work complements several recent papers that seek to reconcile the two biases from different perspectives—most notably, Ba, Bohren, and Imas (2025) and Augenblick, Lazarus, and Thaler (2025).

In the model of Ba, Bohren, and Imas (2025), agents solve belief-updating problems in two stages: mental representation and processing. At the representation stage, attention—channeled

via salience—causes agents to overweight the state most “representative” of the observed signal (Kahneman and Tversky, 1972), generating overreaction. At the processing stage, a second force—cognitive imprecision—induces agents to shade their subjective posteriors toward a cognitive default (Enke and Graeber, 2023), such as an “ignorance prior” that weights all states equally, generating underreaction. The relative strength of these two forces depends on features of the environment, such as the complexity (cardinality) of the state space, leading to overreaction in a complex state space and underreaction in a simple one.

Applying the model of Ba, Bohren, and Imas (2025) to our forecast problem requires specifying how agents mentally represent the problem. We consider two approaches. The first assumes that, when making forecasts, agents adhere to the original state space (Good vs. Bad firms) and use their subjective posteriors over these two states to form forecasts. If this additional step of forming forecasts makes forecast-revision problems more complex,³⁸ Ba, Bohren, and Imas (2025) predicts that agents will be more likely to over-focus on the representative state, leading to greater overreaction in these problems. In the extreme, this mechanism can generate the Exact Representativeness heuristic, but it does not account for the Naive Extrapolation heuristic. The second approach, considered by Ba, Bohren, and Imas (2025) in an extension, assumes that agents channel their attention to the *outcome* most representative of the observed signal. That is, instead of adhering to the original state space of two firm types, agents treat the set of outcomes over which they form beliefs as the relevant state space. This gives rise to a much larger state space and, according to the model, predicts overreaction in the forecast problem as well as an inference-forecast gap.³⁹ Nonetheless, this approach does not directly explain the modal behaviors of Exact Representativeness and Naive Extrapolation. Overall, while complexity may contribute to overreaction in forecast revision, it does not readily account for the specific pattern of modal behaviors we document, nor for how these behaviors vary with similarity manipulations.

While we focus on how different belief-updating problems moderate under- and overreaction, Augenblick, Lazarus, and Thaler (2025) operates within inference problems to uncover an insensitivity to *signal strength*. They show that, consistent with a model in which agents form a noisy estimate of signal strength and shade it toward an intermediate default, individuals overinfer from weak signals and underinfer from strong signals. In our analysis of inference problems, Table A4 shows a similar pattern: the stronger the signals, the less likely participants are to overinfer. However, the same table reveals non-monotonic comparative statics for forecast-revision problems: the

³⁸Here, we adopt a broader notion of complexity which includes not only the size of the state space considered in Ba, Bohren, and Imas (2025) but also other features of the problem, such as the number of computational steps required.

³⁹Consistent with this prediction, Ba, Bohren, and Imas (2025) shows that underreaction dominates in a treatment where the outcome is binarized in the forecast problem. Our paper documents similar evidence: the inference-forecast gap is smaller in *Binary Signal* than in *Baseline*, supporting complexity as at least a partial driver of the gap.

overreaction gap first increases with signal strength—peaking at the Strong bin—before declining and turning to underreaction only at the Strongest signals.⁴⁰ As Table B1 shows, this aggregate pattern can be partly explained by how the prevalence of different simplifying heuristics varies with signal strength in the forecast-revision problem.

6.2 Generalizability of our framework

In Section 5.1, we introduced a framework building on attribute substitution (Kahneman and Frederick, 2002) to account for the inference-forecast gap and the associated emergence of modal behaviors. The central idea is that when decision-makers are unsure of their computed answers, they may substitute more accessible and *similar* statistics as proxy responses. To formalize this process, the framework defines a novel measure of similarity among statistics along three dimensions—statistical measure, essential variable, and conditioning event—allowing us to make predictions about the distributions of answers to statistical problems beyond the current experiment. For example, if the forecast-revision problem elicited the *median* rather than the expectation of the outcome distribution conditional on a signal, the framework would still predict the emergence of the same three modal behaviors. To illustrate, consider Exact Representativeness: under the *Maximum-at-Identity* condition, the similarity between the nontarget statistic $\text{Median}[s_1|\text{representative state}]$ and the target statistic $\text{Median}[s_1|s_0]$ is unchanged relative to the case in which the expectation is elicited. The same reasoning extends to the other two modes.

Our framework can also account for other biases documented in the literature. For example, Aina and Schneider (2025) study an inference problem in which the signal is generated by one of two equally likely models, though the true model is unknown. After observing the signal, participants are first shown the Bayesian posterior over states conditional on each model, $\text{Pr}(\text{state}|\text{signal}, \text{first model})$ and $\text{Pr}(\text{state}|\text{signal}, \text{second model})$. They are then asked to report their unconditional posterior over states, $\text{Pr}(\text{state}|\text{signal})$. The authors find that participants’ responses exhibit two modes: one at the posterior under the more likely model, $\text{Pr}(\text{state}|\text{signal}, \text{more likely model})$, and the other at the prior-weighted average of the two model-specific posteriors, $0.5 \times \text{Pr}(\text{state}|\text{signal}, \text{first model}) + 0.5 \times \text{Pr}(\text{state}|\text{signal}, \text{second model})$. Under our framework, each modal response corresponds to an accessible nontarget statistic. The first differs from the target only in the conditioning event, making it a natural proxy. The second substitutes prior model weights for the correct posterior weights. Since the prior and the posterior share the same statistical measure and essential variable—and differ only in the conditioning event—this substitution is

⁴⁰This result implies that the inference-forecast gap is largest for (moderately) strong signals. Note also that it describes how under- and overreaction vary with signal strength conditional on a DGP, rather than how they vary across DGPs. Thus, it does not conflict with findings that overreaction in forecasts decreases for more persistent time series (Bordalo, Gennaioli, Ma, and Shleifer, 2020; AKLM+, 2023).

plausible under our similarity-based account.⁴¹

Our framework is closely related and complementary to that of BCGK+ (2025), which also explains multi-modal responses in belief-updating tasks. In their model, individuals attend to certain features of the environment while neglecting others, replacing the latter with default values. Beliefs are then formed by correctly aggregating the perceived feature values using Bayes’ rule. Due to the stochastic and discrete nature of attention, this model can explain multi-modality in the distribution of responses. The prevalence of behavioral modes is also shaped by framing, which alters the salience of different features. While both models account for multi-modal responses, our framework differs in ways that make it better suited to the patterns observed in our experiment. We assume that all relevant features are attended to, but that some decision-makers struggle to aggregate them correctly. Instead, they substitute the target statistic with a similar, more accessible nontarget statistic. This similarity-based substitution gives rise to modal behaviors such as Exact Representativeness and Naive Extrapolation—patterns that are difficult to reconcile with correct aggregation of imperfectly perceived features. For example, to generate Naive Extrapolation, agents must attend to the relevant features in a way that makes the Bayesian posterior exactly equal to the value of the realized signal. However, this is unlikely to hold in general, as the realized signal enters Bayes’ rule only through its likelihood ratio, not through its face value. A second difference between the two models lies in the scope: while the model of BCGK+ (2025) applies to settings with binary hypothesis testing, our framework applies more broadly to a wide range of statistical problems, including, but not limited to, expectation-formation problems.

6.3 External validity

This paper documents a disconnect between inference and forecast revision in a controlled setting where the mapping from states to outcomes is deliberately simple. In the field, where fundamentals are more difficult to link to outcomes and frictions are pervasive, this disconnect is likely to be stronger. In corporate finance, for instance, revenue forecasts depend on uncertain factors such as product demand, input costs, and competition, making it challenging to tie forecasts to beliefs about fundamentals. In macroeconomics, the “fundamental state of the economy” is abstract and difficult to quantify, further complicating inference. These complexities make forecasts more likely to diverge from underlying beliefs. Consistent with this view, corporate finance textbooks, such as Welch (2011), recommend revenue-forecasting methods that bypass input forecasts, and empirical evidence shows managers often adopt such approaches (Giustinelli and Rossi, 2023).

In this light, our results may also help to reconcile an apparent disconnect between two strands of studies of belief formation. A long line of experimental work shows that individuals tend to

⁴¹To account precisely for the second mode, our framework would need to be extended to allow substitution of one component in a known formula—in this case, the model weight within the posterior.

underreact to signals when making *inferences* about underlying states (Benjamin, 2019). On the other hand, another literature, using both field and experimental data, shows that individuals tend to *overreact* to recent news when making *forecasts* (e.g., Hey, 1994; Bordalo, Gennaioli, Ma, and Shleifer, 2020), a phenomenon that has been used to explain key anomalies such as excess volatility in financial markets and boom-bust cycles in the macroeconomy (e.g., Bordalo, Gennaioli, Shleifer, and Terry, 2026; Maxted, 2024). Our analysis suggests that the observed gap between these two sets of findings reflects a deeper methodological issue: the framing of belief elicitation—as an inference problem or a forecast problem—fundamentally shapes how individuals process information and update their beliefs.

Our experiment uncovers three heuristics in inference and forecast revision: Naive Extrapolation and Exact Representativeness, which contribute to overreaction in forecast revision; and Non-update, which produces underreaction in both tasks.⁴² While further work is needed to test their external validity (see Appendix F for suggestive evidence) and to identify additional decision rules, several factors suggest these heuristics are likely relevant in the field. First, they rely on cues such as priors, past realizations, and conditional expectations—statistics that are widely available across contexts. Second, they are shaped by psychological forces such as similarity, which also operate in real-world settings. Third, they help individuals manage the greater complexity of field environments. Complexity may also foster each heuristic in distinct ways. Non-update may arise when information is complex or hard to interpret (Enke and Graeber, 2023; Gonçalves, Libgober, and Willis, 2025; Liang, 2025), making priors a natural anchor. Naive Extrapolation may appear when fundamentals are opaque and outcomes resemble a time series; it tends to cause overreaction unless new signals are strong (AKLM+, 2023), as documented in managerial forecasts (Giustinelli and Rossi, 2023) and among professional forecasters (Kohlhas and Walther, 2021). Exact Representativeness—an extreme form of the representativeness heuristic (Kahneman and Tversky, 1972; Bordalo, Gennaioli, and Shleifer, 2018; Bordalo, Gennaioli, Porta, and Shleifer, 2019), in which attention narrows to the representative state—is more likely when integrating across states is cognitively demanding (Martínez-Marquina, Niederle, and Vespa, 2019; Esponda and Vespa, 2023).

Taken together, the inference-forecast gap documented in our experiment may help explain the coexistence of under- and overreaction across field settings. On the one hand, phenomena such as the post-earnings-announcement drift (Bernard and Thomas, 1989) and homeowner underreaction to housing market conditions (Genesove and Mayer, 2001; Case, Shiller, and Thompson, 2012) suggest that beliefs about asset values often underreact to news. On the other hand, forecasts of stock returns and macroeconomic indicators frequently overreact to recent information

⁴²Naive Extrapolation may lead to underreaction when signals are strongly diagnostic or when priors are asymmetric. Nevertheless, under parameters considered in our experiment, this heuristic always leads to overreaction.

(e.g., Greenwood and Shleifer, 2014; Bordalo, Gennaioli, Ma, and Shleifer, 2020), a pattern invoked to explain major anomalies such as excess volatility in financial markets and macroeconomic boom–bust cycles (e.g., Barberis, Greenwood, Jin, and Shleifer, 2015; Bordalo, Gennaioli, Shleifer, and Terry, 2026; Maxted, 2024). Our findings suggest that this discrepancy may arise because people mentally represent some problems as inference tasks and others as forecast-revision tasks, leading them to approach the two in fundamentally different ways. This implies that to understand biased reactions to information in the field, it is essential to first understand how people mentally represent the belief-updating problem they face.

References

- Afrouzi, Hassan, Spencer Y Kwon, Augustin Landier, Yueran Ma, and David Thesmar (2023), “Overreaction in expectations: Evidence and theory.” *The Quarterly Journal of Economics*, 138 (3), 1713–1764.
- Agranov, Marina and Pëllumb Reshidi (2024), “Disentangling suboptimal updating: Complexity, structure, and sequencing.”
- Aina, Chiara and Florian H Schneider (2025), “Weighting competing models.” Technical report.
- Andre, Peter, Carlo Pizzinelli, Christopher Roth, and Johannes Wohlfart (2022), “Subjective models of the macroeconomy: Evidence from experts and representative samples.” *The Review of Economic Studies*, 89 (6), 2958–2991.
- Augenblick, Ned, Eben Lazarus, and Michael Thaler (2025), “Overinference from weak signals and underinference from strong signals.” *The Quarterly Journal of Economics*, 140 (1), 335–401.
- Ba, Cuimin, J Aislinn Bohren, and Alex Imas (2025), “Over- and underreaction to information: Belief updating with cognitive constraints.” *Available at SSRN 4274617*.
- Barberis, Nicholas, Robin Greenwood, Lawrence Jin, and Andrei Shleifer (2015), “X-capm: An extrapolative capital asset pricing model.” *Journal of Financial Economics*, 115 (1), 1–24.
- Barberis, Nicholas, Robin Greenwood, Lawrence Jin, and Andrei Shleifer (2018), “Extrapolation and bubbles.” *Journal of Financial Economics*, 129 (2), 203–227.
- Barrero, Jose Maria (2022), “The micro and macro of managerial beliefs.” *Journal of Financial Economics*, 143 (2), 640–667.

- Benjamin, Daniel J (2019), “Errors in probabilistic reasoning and judgment biases.” *Handbook of Behavioral Economics: Applications and Foundations 1, 2*, 69–186.
- Benjamin, Daniel J, Don A Moore, and Matthew Rabin (2017), “Biased beliefs about random samples: Evidence from two integrated experiments.” Technical report, National Bureau of Economic Research.
- Berk, Jonathan B and Richard C Green (2004), “Mutual fund flows and performance in rational markets.” *Journal of Political Economy*, 112 (6), 1269–1295.
- Bernard, Victor L and Jacob K Thomas (1989), “Post-earnings-announcement drift: delayed price response or risk premium?” *Journal of Accounting research*, 27, 1–36.
- Bland, James R and Yaroslav Rosokha (2021), “Learning under uncertainty with multiple priors: experimental investigation.” *Journal of Risk and Uncertainty*, 62 (2), 157–176.
- Bordalo, Pedro, John Conlon, Nicola Gennaioli, Spencer Kwon, and Andrei Shleifer (2025), “How people use statistics.” *Review of Economic Studies*, rdaf022.
- Bordalo, Pedro, John J Conlon, Nicola Gennaioli, Spencer Y Kwon, and Andrei Shleifer (2023), “Memory and probability.” *The Quarterly Journal of Economics*, 138 (1), 265–311.
- Bordalo, Pedro, Nicola Gennaioli, Yueran Ma, and Andrei Shleifer (2020), “Overreaction in macroeconomic expectations.” *American Economic Review*, 110 (9), 2748–82.
- Bordalo, Pedro, Nicola Gennaioli, Rafael La Porta, and Andrei Shleifer (2019), “Diagnostic expectations and stock returns.” *Journal of Finance*, 74 (6), 2839–2874.
- Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer (2018), “Diagnostic expectations and credit cycles.” *Journal of Finance*, 73 (1), 199–227.
- Bordalo, Pedro, Nicola Gennaioli, Andrei Shleifer, and Stephen J. Terry (2026), “Real credit cycles.” *American Economic Review*, 116 (4), 1274–1308.
- Camerer, Colin F (1987), “Do biases in probability judgment matter in markets? experimental evidence.” *The American Economic Review*, 77 (5), 981–997.
- Caplin, Andrew and Mark Dean (2015), “Revealed preference, rational inattention, and costly information acquisition.” *American Economic Review*, 105 (7), 2183–2203.
- Case, Karl E, Robert J Shiller, and Anne Thompson (2012), “What have they been thinking? home buyer behavior in hot and cold markets.” Technical report, National Bureau of Economic Research.

- Center for Research in Security Prices (CRSP) (n.d.), “CRSP US Stock Databases.” Wharton Research Data Services (WRDS). <https://wrds-www.wharton.upenn.edu/>.
- Chen, Daniel L, Martin Schonger, and Chris Wickens (2016), “oTree—an open-source platform for laboratory, online, and field experiments.” *Journal of Behavioral and Experimental Finance*, 9, 88–97.
- Coutts, Alexander (2019), “Good news and bad news are still news: Experimental evidence on belief updating.” *Experimental Economics*, 22 (2), 369–395.
- Drugowitsch, Jan, Valentin Wyart, Anne-Dominique Devauchelle, and Etienne Koechlin (2016), “Computational precision of mental inference as critical source of human choice suboptimality.” *Neuron*, 92 (6), 1398–1411.
- Ellsberg, Daniel (1961), “Risk, ambiguity, and the savage axioms.” *The quarterly journal of economics*, 75 (4), 643–669.
- Enke, Benjamin (2020), “What you see is all there is.” *Quarterly Journal of Economics*, 135 (3), 1363–1398.
- Enke, Benjamin and Thomas Graeber (2023), “Cognitive uncertainty.” *The Quarterly Journal of Economics*, 138 (4), 2021–2067.
- Enke, Benjamin, Frederik Schwerter, and Florian Zimmermann (2024), “Associative memory, beliefs and market interactions.” *Journal of Financial Economics*, 157, 103853.
- Enke, Benjamin and Florian Zimmermann (2019), “Correlation neglect in belief formation.” *Review of Economic Studies*, 86 (1), 313–332.
- Epstein, Larry G and Yoram Halevy (2024), “Hard-to-interpret signals.” *Journal of the European Economic Association*, 22 (1), 393–427.
- Esponda, Ignacio and Emanuel Vespa (2023), “Contingent thinking and the sure-thing principle: Revisiting classic anomalies in the laboratory.” *Review of Economic Studies*, rdad102.
- Esponda, Ignacio, Emanuel Vespa, and Sevgi Yuksel (2024), “Mental models and learning: The case of base-rate neglect.” *American Economic Review*, 114 (3), 752–782.
- Falk, Armin and Florian Zimmermann (2018), “Information processing and commitment.” *The Economic Journal*, 128 (613), 1983–2002.

- Fan, Tony Q., Yucheng Liang, and Cameron Peng (2026), “Replication package for “The Inference-Forecast Gap in Belief Updating”.” *Econometrica*, data deposited at Zenodo. <https://doi.org/10.5281/zenodo.18928459>.
- Federal Reserve Bank of Philadelphia (n.d.), “Survey of professional forecasters.” Federal Reserve Bank of Philadelphia. <https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/survey-of-professional-forecasters>.
- Fehrler, Sebastian, Baiba Renerte, and Irenaeus Wolff (2023), “Beliefs about others: A striking example of information neglect.”
- Frydman, Cary and Gideon Nave (2017), “Extrapolative beliefs in perceptual and economic decisions: Evidence of a common mechanism.” *Management Science*, 63 (7), 2340–2352.
- Gabaix, Xavier (2014), “A sparsity-based model of bounded rationality.” *Quarterly Journal of Economics*, 129 (4), 1661–1710.
- Genesove, David and Christopher Mayer (2001), “Loss aversion and seller behavior: Evidence from the housing market.” *The quarterly journal of economics*, 116 (4), 1233–1260.
- Gilovich, Thomas, Robert Vallone, and Amos Tversky (1985), “The hot hand in basketball: On the misperception of random sequences.” *Cognitive Psychology*, 17 (3), 295–314.
- Giustinelli, Pamela and Stefano Rossi (2023), “The coherence side of rationality: Rules of thumb, narrow bracketing, and managerial incoherence in corporate forecasts.”
- Glaeser, Edward L and Charles G Nathanson (2017), “An extrapolative model of house price dynamics.” *Journal of Financial Economics*, 126 (1), 147–170.
- Goldstein, William M and Hillel J Einhorn (1987), “Expression theory and the preference reversal phenomena.” *Psychological review*, 94 (2), 236.
- Gonçalves, Duarte, Jonathan Libgober, and Jack Willis (2025), “Retractions: Updating from complex information.” *Review of Economic Studies*, rdaf032.
- Graeber, Thomas (2023), “Inattentive inference.” *Journal of the European Economic Association*, 21 (2), 560–592.
- Greenwood, Robin and Andrei Shleifer (2014), “Expectations of returns and expected returns.” *Review of Financial Studies*, 27 (3), 714–746.

- Hartzmark, Samuel M, Samuel D Hirshman, and Alex Imas (2021), “Ownership, learning, and beliefs.” *The Quarterly journal of economics*, 136 (3), 1665–1717.
- He, Simin and Simas Kučinskas (2024), “Expectation formation with correlated variables.” *The Economic Journal*, 134 (660), 1517–1544.
- Heath, Chip and Amos Tversky (1991), “Preference and belief: Ambiguity and competence in choice under uncertainty.” *Journal of Risk and Uncertainty*, 4 (1), 5–28.
- Hey, John D (1994), “Expectations formation: Rational or adaptive or . . . ?” *Journal of Economic Behavior & Organization*, 25 (3), 329–349.
- Jiang, Zhengyang, Hongqi Liu, Cameron Peng, and Hongjun Yan (2025), “Investor memory and biased beliefs: Evidence from the field.” *The Quarterly Journal of Economics*, 140 (4), 2749–2804.
- Kahana, Michael Jacob (2012), *Foundations of human memory*, OUP USA.
- Kahneman, Daniel and Shane Frederick (2002), “Representativeness revisited: Attribute substitution in intuitive judgment.” *Heuristics and biases: The psychology of intuitive judgment*, 49, 81.
- Kahneman, Daniel and Amos Tversky (1972), “Subjective probability: A judgment of representativeness.” *Cognitive psychology*, 3 (3), 430–454.
- Kendall, Chad and Ryan Oprea (2024), “On the complexity of forming mental models.” *Quantitative Economics*, 15 (1), 175–211.
- Kieren, Pascal, Jan Müller-Dethard, and Martin Weber (2023), “Disconfirming information and overreaction in expectations: Evidence from the lab and field.” *Working paper*.
- Kohlhas, Alexandre N and Ansgar Walther (2021), “Asymmetric attention.” *American Economic Review*, 111 (9), 2879–2925.
- Liang, Yucheng (2025), “Learning from unknown information sources.” *Management Science*, 71 (5), 3873–3890.
- Liao, Jingchi, Cameron Peng, and Ning Zhu (2022), “Extrapolative bubbles and trading volume.” *The Review of Financial Studies*, 35 (4), 1682–1722.
- Logan, Gordon D (2021), “Serial order in perception, memory, and action.” *Psychological Review*, 128 (1), 1.

- Martínez-Marquina, Alejandro, Muriel Niederle, and Emanuel Vespa (2019), “Failures in contingent reasoning: The role of uncertainty.” *American Economic Review*, 109 (10), 3437–3474.
- Maxted, Peter (2024), “A macro-finance model with sentiment.” *Review of Economic Studies*, 91 (1), 438–475.
- Moreno, Othon M and Yaroslav Rosokha (2016), “Learning under compound risk vs. learning under ambiguity-an experiment.” *Journal of Risk and Uncertainty*, 137–162.
- Mullainathan, Sendhil (2002), “Thinking through categories.” *Working paper*.
- Mullainathan, Sendhil, Joshua Schwartzstein, and Andrei Shleifer (2008), “Coarse thinking and persuasion.” *Quarterly Journal of Economics*, 123 (2), 577–619.
- Nielsen, Kirby (2020), “Preferences for the resolution of uncertainty and the timing of information.” *Journal of Economic Theory*, 189, 105090.
- Palan, Stefan and Christian Schitter (2018), “Prolific.ac—a subject pool for online experiments.” *Journal of Behavioral and Experimental Finance*, 17, 22–27.
- Rabin, Matthew and Dimitri Vayanos (2010), “The gambler’s and hot-hand fallacies: Theory and applications.” *Review of Economic Studies*, 77 (2), 730–778.
- Rothbart, Myron and Mark Snyder (1970), “Confidence in the prediction and postdiction of an uncertain outcome.” *Canadian Journal of Behavioural Science*, 2 (1), 38.
- Shah, Anuj K and Daniel M Oppenheimer (2008), “Heuristics made easy: an effort-reduction framework.” *Psychological Bulletin*, 134 (2), 207.
- Sims, Christopher A (2003), “Implications of rational inattention.” *Journal of monetary Economics*, 50 (3), 665–690.
- Suetens, Sigrid, Claus B Galbo-Jørgensen, and Jean-Robert Tyran (2016), “Predicting lotto numbers: a natural experiment on the gambler’s fallacy and the hot-hand fallacy.” *Journal of the European Economic Association*, 14 (3), 584–607.
- UBS/Gallup (n.d.), “UBS/Gallup Index of Investor Optimism.” UBS and Gallup, archived at the Roper Center for Public Opinion Research. <https://ropercenter.cornell.edu/ubs-index-investor-optimism-1996-2007>.
- Welch, Ivo (2011), *Corporate finance*, Ivo Welch.
- Woodford, Michael (2020), “Modeling imprecision in perception, valuation, and choice.” *Annual Review of Economics*, 12, 579–601.

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 The subsample of participants who answered comprehension questions correctly in one pass

We start by examining the inference-forecast gap in a subsample of the *Baseline* participants who answered all our comprehension questions correctly in one pass.⁴³ This subsample includes 79 (28.3%) of the 279 participants in *Baseline*.⁴⁴

Table A1: Aggregate patterns in *Baseline*: subsample who answered all comprehension questions correctly in one pass

<i>N</i> =79, Obs.=605	Classification			Update
	Underreaction	Near-rational	Overreaction	Mean (s.e.)
<i>Inference</i>	50.1%	24%	26%	17.7 (1.2)
<i>Forecast Revision</i>	38%	10.6%	51.4%	28.3 (2.4)
Rational				23.4 (.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.

⁴³Participants can make multiple attempts at the comprehension questions, and as long as they correctly answer the questions on their last attempt, they are allowed to proceed to the next stage of the experiment. Nonetheless, most of our comprehension questions are not multiple-choice questions but in the style of questions-and-answers, where subjects must fill in the numerical answers by themselves. Therefore, participants could not pass the comprehension check simply by using trial and error on all the possible options without reading the instructions. Instead, they need to rely on the instructions—most notably the descriptions of features of the belief-updating problems—to answer all the questions correctly.

⁴⁴The fraction of retained participants may appear low, as one of our comprehension questions is a rather challenging one related to the gambler’s fallacy. If we ignore this particular question, the fraction of participants who answered all questions correctly in one pass goes up to 68.8% (192/279). Using this more “lenient” subsample again produces very similar results (available upon request).

Table A1 shows the results in this subsample, which are very similar to results from the full sample: On average, there is significant overreaction in *Forecast Revision* and underreaction in *Inference*. The gap in updates between these two parts is significant, as is shown in a regression analysis in Column (2) of Table A7.

A.2 A more “reasonable” subsample based on posterior beliefs

We now examine 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 A2: Aggregate patterns in *Baseline*: subsample with “reasonable” updates

<i>N</i> =269, Obs.=1366	Classification			Update
	Underreaction	Near-rational	Overreaction	Mean (s.e.)
<i>Inference</i>	54.5%	17.9%	27.7%	17.7 (.9)
<i>Forecast Revision</i>	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 A2 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*.⁴⁵ The gap in updates between these two parts is significant, as is shown in a regression analysis in Column (3) of Table A7.

A.3 Priors and signals

The inference-forecast gap exists in all the eight problems with different DGPs (see Table A3). Notably, the eight problems include DGPs with symmetric and asymmetric priors, indicating that

⁴⁵Further restricting the sample to participants who answered all the comprehension questions correctly also produces very similar results (available upon request).

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 A4 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 A5 and in Column (4) of Table A7, is similar: there is more overreaction and less underreaction in *Forecast Revision* than in *Inference*.

A.4 Order between parts

The gap is also robust to different ordering of the five parts. Table A6 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.5 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 A9 shows 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 a substantially smaller inference-forecast gap than that observed for 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.

A.6 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

Table A3: Aggregate patterns in *Baseline* (by problem)

		Classification			Update
		Underreaction	Near-rational	Overreaction	Mean (s.e.)
$\Pr(G) = 50\%$	<i>Inference</i>	71.1%	19.8%	9.2%	18.7 (1.2)
$\sigma = 50$	<i>Forecast Revision</i>	44.3%	12.1%	43.6%	31.2 (2.4)
(Obs. = 273)	Rational				35.9 (.8)
$\Pr(G) = 50\%$	<i>Inference</i>	68.2%	16.9%	14.9%	17.0 (1.2)
$\sigma = 60$	<i>Forecast Revision</i>	47.9%	6.5%	45.6%	28.6 (2.8)
(Obs. = 261)	Rational				31.8 (.8)
$\Pr(G) = 50\%$	<i>Inference</i>	64.8%	13.5%	21.7%	15.3 (1.1)
$\sigma = 70$	<i>Forecast Revision</i>	40.8%	7.1%	52.1%	29 (2.6)
(Obs. = 267)	Rational				27 (.8)
$\Pr(G) = 50\%$	<i>Inference</i>	64.7%	12.6%	22.7%	13.9 (1.2)
$\sigma = 80$	<i>Forecast Revision</i>	40.9%	4.5%	54.6%	34.1 (3.6)
(Obs. = 269)	Rational				25 (.8)
$\Pr(G) = 50\%$	<i>Inference</i>	50.6%	18.4%	31.1%	16.2 (1.1)
$\sigma = 90$	<i>Forecast Revision</i>	36.7%	4.1%	59.2%	37.3 (3.3)
(Obs. = 267)	Rational				21.8 (.7)
$\Pr(G) = 50\%$	<i>Inference</i>	51.3%	16.1%	32.6%	13.1 (1.2)
$\sigma = 100$	<i>Forecast Revision</i>	32.2%	8.2%	59.6%	38.3 (3.5)
(Obs. = 267)	Rational				19.7 (.7)
$\Pr(G) = 80\%$	<i>Inference</i>	57.4%	13.3%	29.3%	10.6 (1.3)
$\sigma = 100$	<i>Forecast Revision</i>	38.1%	3%	58.9%	34.1 (4.1)
(Obs. = 270)	Rational				12.8 (.6)
$\Pr(G) = 20\%$	<i>Inference</i>	58.1%	10.7%	31.1%	10 (1.5)
$\sigma = 100$	<i>Forecast Revision</i>	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 A4: Aggregate patterns in *Baseline* (by signal strength)

Signal Strength		Classification			Update
		Underreaction	Near-rational	Overreaction	Mean (s.e.)
Weakest (Obs. = 239)	<i>Inference</i>	47.7%	23.0%	29.3%	4.5 (.8)
	<i>Forecast Revision</i>	48.1%	11.7%	40.2%	10.3 (1.5)
	Rational				6.5 (.2)
Weak (Obs. = 313)	<i>Inference</i>	59.4%	13.7%	26.8%	9.6 (.9)
	<i>Forecast Revision</i>	44.4%	5.4%	50.2%	19.1 (2.1)
	Rational				15.9 (.2)
Medium (Obs. = 280)	<i>Inference</i>	63.9%	10.4%	25.7%	15.1 (1.1)
	<i>Forecast Revision</i>	37.5%	5.0%	57.5%	33.4 (2.7)
	Rational				25.1 (.1)
Strong (Obs. = 300)	<i>Inference</i>	65.0%	12.3%	22.7%	20.4 (1.4)
	<i>Forecast Revision</i>	34.7%	4.0%	61.3%	49.6 (4.1)
	Rational				34.4 (.2)
Strongest (Obs. = 362)	<i>Inference</i>	63.8%	25.1%	11.0%	25.8 (1.4)
	<i>Forecast Revision</i>	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 A5: Aggregate patterns in *Baseline*: subsample with correct priors

<i>N</i> =279, Obs.=1502	Classification			Update
	Underreaction	Near-rational	Overreaction	Mean (s.e.)
<i>Inference</i>	57.7%	17.9%	24.4%	15.7 (.8)
<i>Forecast Revision</i>	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 A6: Aggregate patterns in *Baseline* (by order between parts)

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

revenue growth (rather than stock price growth); we find a quantitatively smaller but still significant gap with this alternative framing. Table A10 shows these results in regressions.

A.7 Regression analyses

Table A7: The inference-forecast gap in *Baseline* under various sample restrictions.

	Update			
	Full Sample	Best comprehension	“Reasonable” updates	Correct priors
	(1)	(2)	(3)	(4)
Forecast Revision	18.385 (2.279)	10.594 (2.578)	6.682 (1.210)	11.751 (2.530)
Rational Update	1.035 (0.069)	0.817 (0.104)	0.578 (0.041)	0.926 (0.074)
Problem FE	Yes	Yes	Yes	Yes
Subject FE	Yes	Yes	Yes	Yes
Observations	4288	1210	2732	3004
R^2	0.314	0.359	0.463	0.341

Notes: 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 exclude participants who did not answer all the comprehension questions correctly in one pass. In Column (3), 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 (4), based on the full sample, we further drop observations with an incorrect answer for either *Inference Prior* or *Forecast Prior*.

Table A8: The inference-forecast gap in *Baseline* excluding modal behaviors

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

Notes: 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 A9: Heterogeneity of the inference-forecast gap across demographics

	Update
<i>Forecast Revision</i>	30.942 (3.849)
Male \times <i>Forecast Revision</i>	-5.152 (4.544)
College \times <i>Forecast Revision</i>	-2.804 (4.504)
Investor \times <i>Forecast Revision</i>	-2.310 (4.516)
Familiar with Stats \times <i>Forecast Revision</i>	-6.481 (5.080)
Familiar with Econ \times <i>Forecast Revision</i>	-6.117 (5.436)
High Comprehension \times <i>Forecast Revision</i>	-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: 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 A10: Heterogeneity of the inference-forecast gap across alternative framing

	Update
Stock Price \times <i>Forecast Revision</i>	21.187 (3.000)
Revenue \times <i>Forecast Revision</i>	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: 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 A11: The inference-forecast gap across different treatments

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

Notes: 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 A12: The inference-forecast gap in *Binary Signal* treatment

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

Notes: Standard errors are clustered by participant. 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.

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 Modes of behavior by signal strength

For the subsample with symmetric (objective) priors, we further examine how the modes of behavior 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 B1 shows the results. In *Forecast Revision*, the incidence of Exact Representativeness increases monotonically by signal strength, which can at least partly explain the increase in forecast overreaction as signals become stronger. It is less clear-cut how the incidence of Naive Extrapolation varies by signal strength. In both *Inference* and *Forecast Revision*, the prevalence of Non-update generally decreases as signals become stronger.

Table B1: Modes of behavior in *Baseline* (by signal strength)

Signal Strength		Modes of Behavior		
		Non-update	Exact Representativeness	Naive Extrapolation
Weakest (Obs. = 239)	<i>Inference</i>	36.8%	0.4%	12.1%
	<i>Forecast Revision</i>	38.5%	7.9%	15.5%
Weak (Obs. = 313)	<i>Inference</i>	33.5%	1.9%	4.8%
	<i>Forecast Revision</i>	24%	21.7%	18.5%
Medium (Obs. = 280)	<i>Inference</i>	30.4%	1.4%	0.7%
	<i>Forecast Revision</i>	23.2%	26.4%	11.4%
Strong (Obs. = 300)	<i>Inference</i>	26%	2.7%	1.3%
	<i>Forecast Revision</i>	18%	26.7%	13.3%
Strongest (Obs. = 362)	<i>Inference</i>	27.6%	5%	0%
	<i>Forecast Revision</i>	22.4%	26.8%	4.7%

Notes: The columns present the percentages of answers classified into the modes of Non-update, Exact Representativeness, and Naive Extrapolation, separately by *Inference* and *Forecast Revision*. 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 $[40, 50]$.

B.2 Problems with asymmetric priors

Table B2 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-update is prevalent in both *Inference* and *Forecast Revision*, while Exact Representativeness and Naive Extrapolation show up almost exclusively in the latter.

Table B2: Modes of behavior in *Baseline*: subsample with asymmetric priors

Mode	Criterion for answer	<i>Inference</i>	<i>Forecast Revision</i>
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 rep-

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.3 Relaxing criteria for classification

Table B3 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.4 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 B4 shows the joint distribution of types across the two parts. The number of participants who can be classified into at least one mode is 72 in *Inference*⁴⁶ and 105 in *Forecast Revision*, 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.5 Modes of behavior in other treatments

Table B5 presents results on the modal behaviors in *Deterministic Outcome*. The distribution of modes is similar to *Baseline*. Non-update is prevalent in both *Inference* and *Forecast Revision*, while Exact Representativeness and Naive Extrapolation are only prevalent in the latter.

Table B6 shows that the distribution of modal behaviors in *Binary Signal* are also similar to those in *Baseline*. Non-update is prevalent in both *Inference* and *Forecast Revision*. In *Forecast Revision*, 17.4% of the answers equal the outcome probability of the representative state, which constitutes the behavioral mode of Exact Representativeness. Very few answers are classified as Exact Representativeness in *Inference*.

Table B7 presents the distribution of modal behaviors in *Nudge*. The fraction of Non-update in *Inference* is 53.2%, a notable increase from the 29.7% in *Baseline*. However, the fraction of Non-update in *Forecast Revision* remains roughly the same as in *Baseline*, as does the fraction of

⁴⁶One participant is classified into both Exact Representativeness and Naive Extrapolation in *Inference*.

Table B3: Modes of behavior in *Baseline* with relaxed criteria for mode classification

Mode	Criterion for answer	<i>Inference</i>	<i>Forecast Revision</i>
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.

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 B4: Joint distribution of *Inference* types and *Forecast Revision* types in *Baseline*

<i>Inference</i> type \ <i>Forecast Revision</i> 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 B5: Modes of behavior in *Deterministic Outcome*

Mode	Criterion for answer	<i>Inference</i>	<i>Forecast Revision</i>
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 B6: 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%
	Non-update	$\Pr(\theta s_0) = Pr(\theta)$	27.1%
<i>Inference</i>	Exact Representativeness	$\Pr(G s_0) = 100\%$ if $s_0 = \text{up}$	3.1%
		$\Pr(G s_0) = 0$ if $s_0 = \text{down}$	
	Unclassified		67.6%
<i>Forecast Revision</i>	Non-update	$\Pr(s_1 s_0) = Pr(s_1)$	19.8%
	Exact Representativeness	$\Pr(s_1 s_0) = Pr(s_1 G)$ if $s_0 = \text{up}$	17.4%
		$\Pr(s_1 s_0) = Pr(s_1 B)$ if $s_0 = \text{down}$	
	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 B7: Modes of behavior in *Nudge*

Mode	Criterion for answer	<i>Inference</i>	<i>Forecast Revision</i>
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.

B.6 Proof for Prediction 1

Prediction. Assume that

$$\mu_{\hat{y}}^I \geq \mu_{\hat{y}}^{FR}.$$

Define the proportional difference in target-nontarget similarity between Inference and Forecast Revision as

$$Diff_{NE} = \frac{\Delta(x_{NE}, y^{FR}) - \Delta(x_{NE}, y^I)}{\Delta(x_{NE}, y^I)}, \quad Diff_{ER} = \frac{\Delta(x_{ER}, y^{FR}) - \Delta(x_{ER}, y^I)}{\Delta(x_{ER}, y^I)}.$$

and the tolerance ratios as

$$R_{ER} = \frac{\mu_{\hat{y}}^I + \Delta(x_{NU}, y^I)}{\Delta(x_{ER}, y^I)}, \quad R_{NE} = \frac{\mu_{\hat{y}}^I + \Delta(x_{NU}, y^I)}{\Delta(x_{NE}, y^I)}.$$

If the proportional increases satisfy

$$\frac{1}{1 + R_{ER}} < \frac{Diff_{NE}}{Diff_{ER}} < 1 + R_{NE},$$

then

$$\pi(x_{ER}, y^I) < \pi(x_{ER}, y^{FR}), \quad \pi(x_{NE}, y^I) < \pi(x_{NE}, y^{FR}).$$

Proof. The prediction that the frequency of Exact Representativeness is higher in Forecast Revision is equivalent to

$$\begin{aligned} & \frac{\Delta(x_{ER}, y^I)}{\mu_{\hat{y}}^I + \Delta(x_{NU}, y^I) + \Delta(x_{ER}, y^I) + \Delta(x_{NE}, y^I)} < \frac{\Delta(x_{ER}, y^{FR})}{\mu_{\hat{y}}^{FR} + \Delta(x_{NU}, y^{FR}) + \Delta(x_{ER}, y^{FR}) + \Delta(x_{NE}, y^{FR})} \\ \iff & \frac{\Delta(x_{ER}, y^I)}{\mu_{\hat{y}}^I + \Delta(x_{NU}, y^I) + \Delta(x_{ER}, y^I) + \Delta(x_{NE}, y^I)} < \frac{\Delta(x_{ER}, y^{FR})}{\mu_{\hat{y}}^I + \Delta(x_{NU}, y^{FR}) + \Delta(x_{ER}, y^{FR}) + \Delta(x_{NE}, y^{FR})} \\ \iff & \frac{\mu_{\hat{y}}^I + \Delta(x_{NU}, y^{FR}) + \Delta(x_{ER}, y^{FR}) + \Delta(x_{NE}, y^{FR})}{\mu_{\hat{y}}^I + \Delta(x_{NU}, y^I) + \Delta(x_{ER}, y^I) + \Delta(x_{NE}, y^I)} < \frac{\Delta(x_{ER}, y^{FR})}{\Delta(x_{ER}, y^I)} \\ \iff & 1 + \frac{\Delta(x_{ER}, y^{FR}) - \Delta(x_{ER}, y^I) + \Delta(x_{NE}, y^{FR}) - \Delta(x_{NE}, y^I)}{\mu_{\hat{y}}^I + \Delta(x_{NU}, y^I) + \Delta(x_{ER}, y^I) + \Delta(x_{NE}, y^I)} < 1 + \frac{\Delta(x_{ER}, y^{FR}) - \Delta(x_{ER}, y^I)}{\Delta(x_{ER}, y^I)} \\ \iff & \frac{\Delta(x_{ER}, y^{FR}) - \Delta(x_{ER}, y^I) + \Delta(x_{NE}, y^{FR}) - \Delta(x_{NE}, y^I)}{\Delta(x_{ER}, y^{FR}) - \Delta(x_{ER}, y^I)} < \frac{\mu_{\hat{y}}^I + \Delta(x_{NU}, y^I) + \Delta(x_{ER}, y^I) + \Delta(x_{NE}, y^I)}{\Delta(x_{ER}, y^I)} \\ \iff & \frac{\Delta(x_{NE}, y^{FR}) - \Delta(x_{NE}, y^I)}{\Delta(x_{ER}, y^{FR}) - \Delta(x_{ER}, y^I)} < \frac{\mu_{\hat{y}}^I + \Delta(x_{NU}, y^I) + \Delta(x_{NE}, y^I)}{\Delta(x_{ER}, y^I)} \\ \iff & \frac{\frac{\Delta(x_{NE}, y^{FR}) - \Delta(x_{NE}, y^I)}{\Delta(x_{ER}, y^{FR}) - \Delta(x_{ER}, y^I)}}{\frac{\Delta(x_{ER}, y^{FR}) - \Delta(x_{ER}, y^I)}{\Delta(x_{ER}, y^I)}} < \frac{\mu_{\hat{y}}^I + \Delta(x_{NU}, y^I) + \Delta(x_{NE}, y^I)}{\Delta(x_{NE}, y^I)} = 1 + \frac{\mu_{\hat{y}}^I + \Delta(x_{NU}, y^I)}{\Delta(x_{NE}, y^I)}. \end{aligned}$$

Similarly, the condition

$$\frac{1}{1 + R_{ER}} < \frac{\text{Diff}_{NE}}{\text{Diff}_{ER}}$$

leads to the prediction that the frequency of Naive Extrapolation is higher in *Forecast Revision*. \square

C Beliefs without a 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*.

C.1 Evidence from *Inference Prior* and *Forecast Prior*

Figure C1 shows the distribution of answers in *Inference Prior* and *Forecast Prior* in the *Baseline* treatment. Figure C2 shows the same results in the *Nudge* treatment. In both treatments, 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.

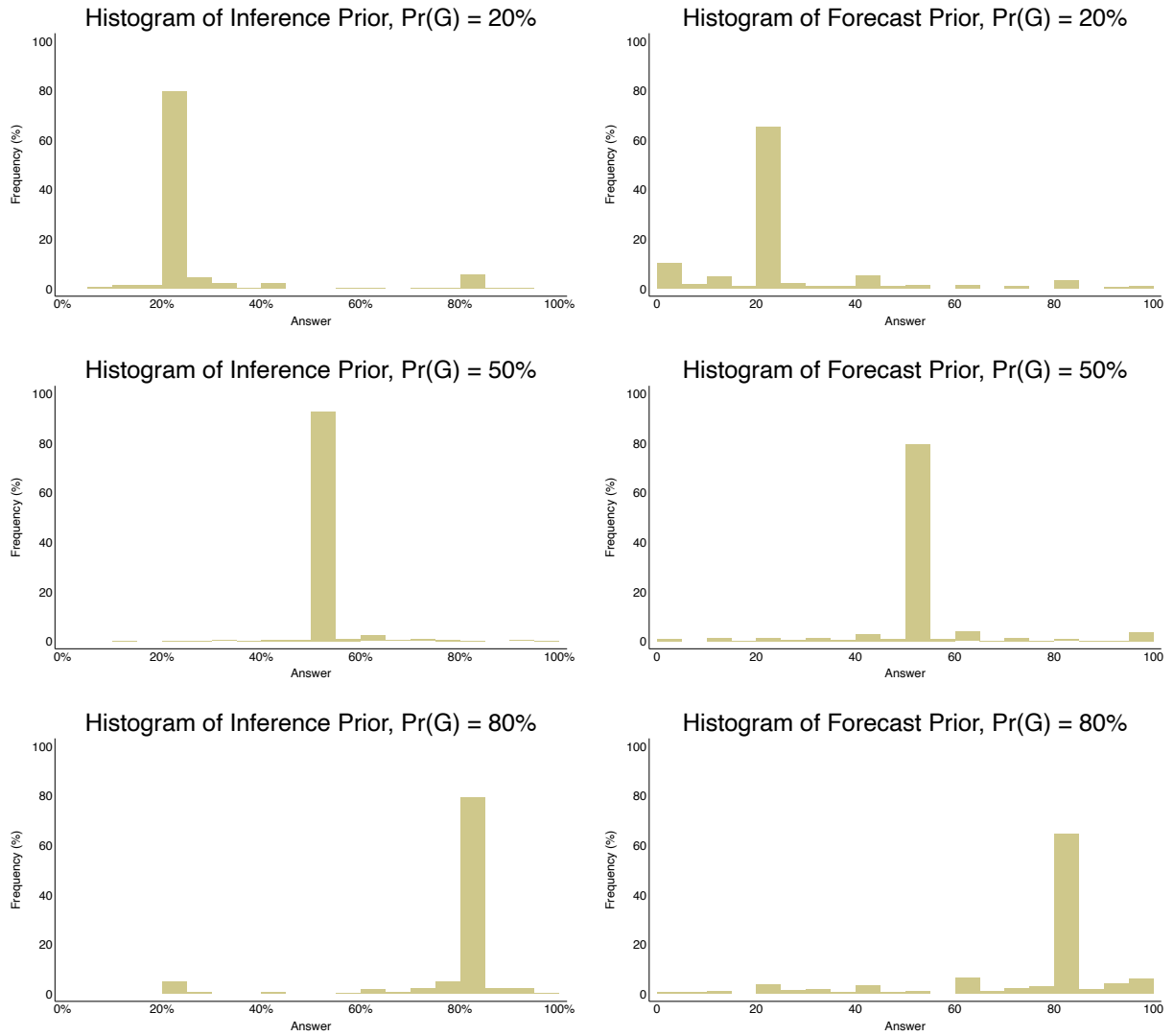


Figure C1: Distributions of answers in *Inference Prior* and *Forecast Prior* in *Baseline*

Notes: We winsorize the answers in *Forecast Prior* at 0 and 100 as in the figures.

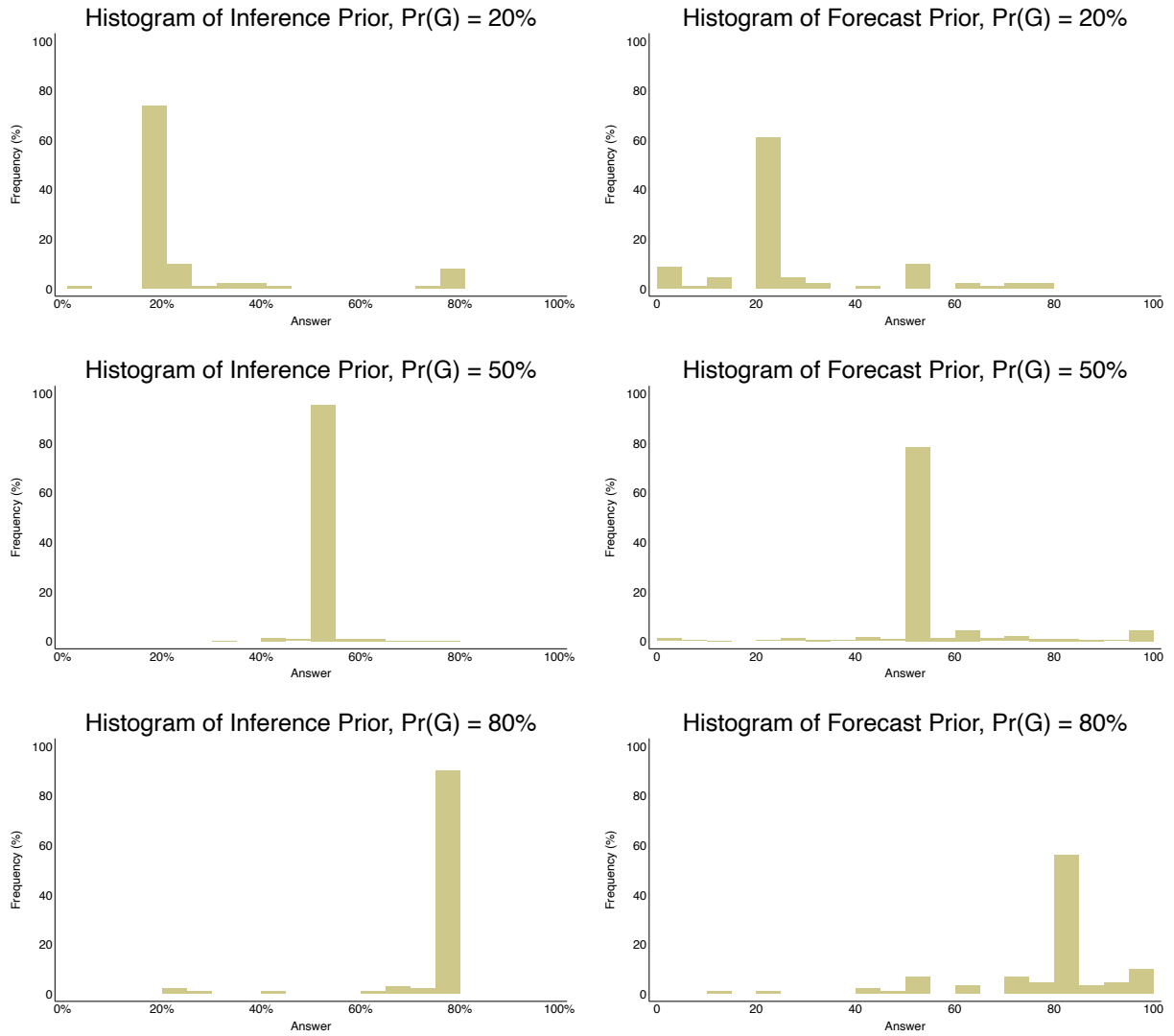


Figure C2: Distributions of answers in *Inference Prior* and *Forecast Prior* in *Nudge*

Notes: We winsorize the answers in *Forecast Prior* at 0 and 100 as in the figures.

C.2 Evidence from *Expectation Formation*

Like *Forecast Prior*, the *Expectation Formation* part asks about participants to report their expectations of the outcome without seeing any realized signal. Unlike *Forecast Prior*, in *Expectation Formation*, the distribution over states in each problem is set to match the posterior reported by the same participant in the corresponding inference problem. In this section, we examine the *Expectation Formation* data from both *Nudge* and *Baseline* treatments to assess whether errors in these problems can explain the degree of overreaction in *Forecast Revision* problems.

C.2.1 *Nudge* treatment

Earlier, Figure 3 showed how much expectation-formation answers deviate from the correct answers prescribed by the LoIE in the *Nudge* treatment. Deviations in the direction of amplifying reactions are generally small and, when analyzed one prior-bin at a time, statistically insignificant. To quantify the degree of such deviations relative to the overreaction observed in *Forecast Revision*, we first consider the following one-parameter distortion function in Goldstein and Einhorn (1987):⁴⁷

$$\text{Expectation Formation Answer} = 100 \times \frac{p^\gamma}{p^\gamma + (1-p)^\gamma} + \epsilon, \quad (11)$$

where p is the prior in the expectation-formation problem (probability of the Good state) and γ governs the degree of distortion. When $\gamma = 1$, expectation-formation answers, on average, do not deviate from the correct answers. With a larger γ , expectation-formation answers increasingly deviate from the correct answers in the direction of amplifying reactions, producing an “S-shaped” distortion. We estimate equation (11) using data from *Nudge* with non-linear least-squares, and obtain an estimate of $\hat{\gamma} = 1.79$ (s.e. = 0.26). Figure C3 shows the fitted curve as well as a binned scatterplot.

We then consider the question of whether the degree of overreaction in *Forecast Revision* can be matched with a larger γ , i.e., by amplifying the degree of distortion in expectation-formation. Specifically, we consider agents who distort their own *Inference* answer using the same process to arrive at their *Forecast Revision* answer:

$$\text{Forecast Revision Answer} = 100 \times \frac{p^\gamma}{p^\gamma + (1-p)^\gamma}, \quad (12)$$

where p is their own *Inference* answer. Even when γ is set to be infinitely large, the predicted fraction of overreactions in *Forecast Revision* is 46.8%, below the actual fraction of 52.3%. Similarly,

⁴⁷The original function in Goldstein and Einhorn (1987) includes an additional δ parameter in front of p^γ that governs elevation. Here, we abstract away from δ by setting it to 1, as including it does not improve the explanatory power of the function for our data.

the predicted average update is only 18.1, smaller than the actual average of 29.8.

Why are distortions in the use of *Inference* answers in expectation-formation not sufficient to generate the overreaction patterns in *Forecast Revision*? The main reason is that a large number of participants report exactly 50% as their *Inference* answer—a modal behavior we refer to as Non-update (also see Table B7). In such cases, equation (12) implies a *Forecast Revision* answer of 50 regardless of γ .⁴⁸ In other words, a posterior of 50% over the two states cannot be systematically distorted in the signal-consistent direction through standalone expectation-formation problems (without a direct influence of the signal).

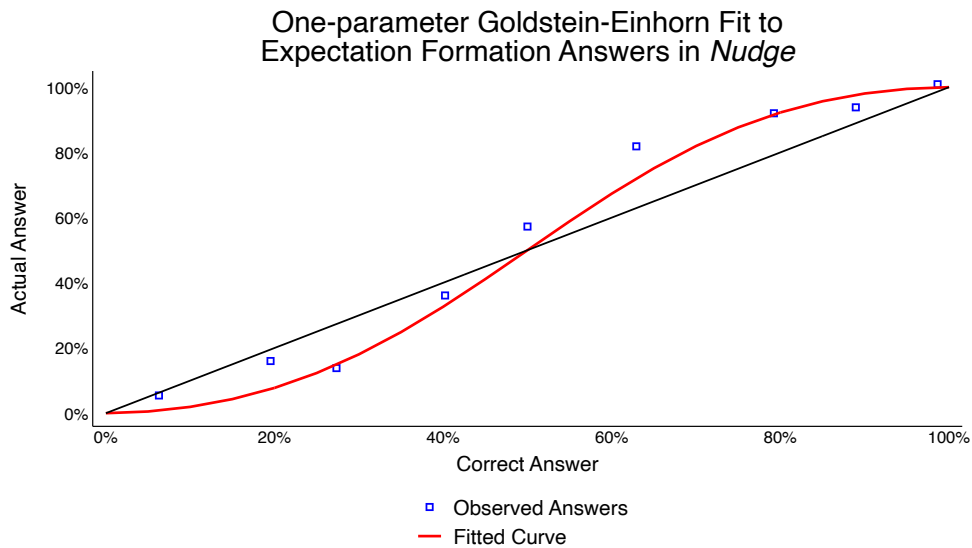


Figure C3: Deviations from LoIE in expectation-formation problems in *Nudge*

Notes: The hollow squares are binned scatterplots of the expectation-formation data in *Nudge*. The red curve is the best-fitting one-parameter distortion function in equation (11) (Goldstein and Einhorn, 1987).

⁴⁸This result does not change even if we add the elevation parameter δ in Goldstein and Einhorn (1987), since such an elevation cannot be conditioned on the direction of the signal.

C.2.2 *Baseline* treatment

Figure C4 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 fail to account for much of the inference-forecast gap. To further quantify the degree of such errors relative to the overreaction in *Forecast Revision*, we estimate equation (11) on the *Expectation Formation* data from *Baseline* using non-linear least-squares, and obtain an estimate of $\hat{\gamma} = 1.49$ (s.e.= 0.06). Thus, the magnitude of expectation-formation errors in *Baseline* is somewhat smaller than in *Nudge*. Figure C3 shows the fitted curve as well as a binned scatterplot of the *Expectation Formation* data from *Baseline*.

As before, we then examine whether the degree of overreaction in *Forecast Revision* can be matched by amplifying the degree of distortion in expectation-formation captured by γ . We consider agents who distort their own *Inference* answer in the same process to arrive at their *Forecast Revision* answer:

$$\text{Forecast Revision Answer} = 100 \times \frac{p^\gamma}{p^\gamma + (1-p)^\gamma}, \quad (13)$$

where p is their own *Inference* answer. If $\gamma = 3.06$, equation (13) can match the fact that 53.9% of *Forecast Revision* answers in *Baseline* are classified as overreaction. However, even when γ is set to be infinitely large, the model predicts an average update of only 25.7 in *Forecast Revision*, falling short of the actual average of 32.7. This gap arises in part because, as in the *Nudge* treatment, many participants in *Baseline* report an *Inference* answer of exactly 50% (see Table 10). For these participants, equation (12) always yields a *Forecast Revision* answer of 50, regardless of how large γ is. These flat posteriors over the states constrain the model's ability to generate large updates, limiting the extent to which distortion in expectation-formation alone can explain the observed overreaction in *Forecast Revision*.

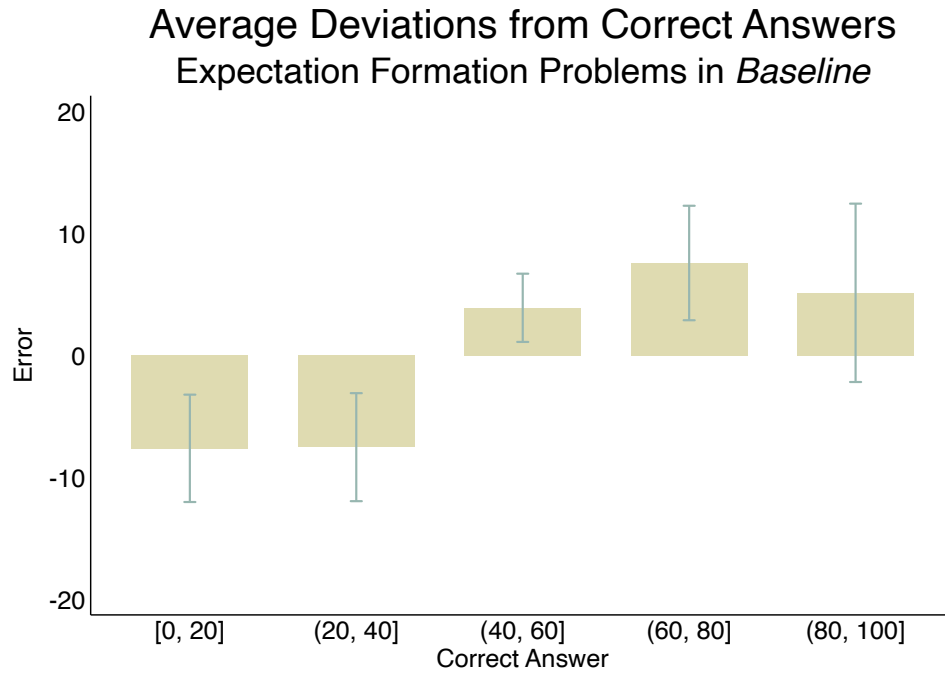


Figure C4: Deviations from LoIE in expectation-formation problems in *Baseline*

Notes: Standard errors are clustered by participant.

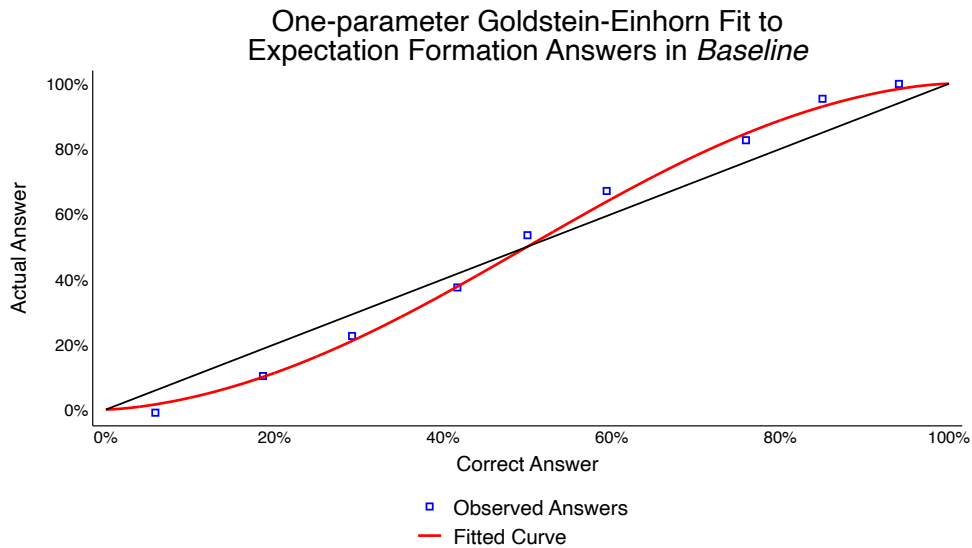


Figure C5: Deviations from LoIE in expectation-formation problems in *Baseline*

Notes: The hollow squares are binned scatterplots of the expectation-formation data in *Baseline*. The red curve is the best-fitting one-parameter distortion function in equation (11) (Goldstein and Einhorn, 1987).

D Additional evidence on timing as a complementary mechanism

In this section, we explore the timing of realization of the elicited statistic as a complementary mechanism for the inference-forecast gap. Note that, in all our treatments in the paper, 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 conjecture builds on several earlier papers that document a timing effect in decision-making under uncertainty. For example, 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, Moore, and Rabin (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.

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. 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.⁴⁹ 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 in either condition, the rational benchmarks for corresponding inference and forecast/precast-revision problems are always exactly the same.

In Table D1, 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 overreac-

⁴⁹To 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.

tion 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; $p = 0.01$). Table D2 confirms, in a regression framework, that the inference-precast gap in *Past* is statistically significantly smaller than the inference-forecast gap in *Future*.

Table D1: Aggregate patterns in *Timing*

<i>Future</i> Condition	Classification			Update
	Underreaction	Near-rational	Overreaction	Mean (s.e.)
N=61, Obs.=470				
<i>Inference</i>	62.3%	12.8%	24.9%	14.0 (1.3)
<i>Forecast Revision</i>	30.0%	6.0%	64.0%	44.9 (5.1)
Rational				23.3 (.6)
<i>Past</i> Condition	Classification			Update
Underreaction	Near-rational	Overreaction	Mean (s.e.)	
N=59, Obs.=458				
<i>Inference</i>	65.7%	16.2%	18.1%	13.0 (1.4)
“Precast” <i>Revision</i>	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 D3 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 this section, we do provide 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)

Table D2: The inference-forecast gap across the two conditions in the *Timing* treatment.

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

Notes: 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.

or a preference for commitment (Falk and Zimmermann, 2018). This conjecture may 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 conjecture may also speak to the asymmetry between forward- and backward-looking gambler's fallacy in Benjamin, Moore, and Rabin (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.

Table D3: Modes of behavior in *Timing*

Mode	<i>Future</i> Condition		<i>Past</i> Condition	
	<i>Inference</i>	<i>Forecast Revision</i>	<i>Inference</i>	<i>“Precast” Revision</i>
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%	49.1%	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.

E Evidence from the *Less Similar* treatment

In a second treatment to test Prediction 2 of the similarity-based framework, which we call *Less Similar*, we reframe the forecast-revision problem to *decrease* the similarity between its target statistic and the overreaction-inducing nontarget statistics. In this treatment, the state variable, the signal, and the inference problem remain the same as in *Baseline*. We modify the forecast-revision problem as follows. After observing the realized stock price growth, participants are asked about *the probability that the firm’s revenue goes 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.

In this treatment, the heuristic of Exact Representativeness can arise if participants use one of the following two nontarget statistics with values of either 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 *Baseline*, the first nontarget statistic $\mathbb{E}[\text{price}|\text{representative state}]$ has now become less similar to the target statistic in *Forecast Revision*— $\Pr(\text{revenue up}|\text{realized price})$ —as the latter is a probability distribution over revenue movements. Although the second nontarget statistic $\Pr(\text{revenue up}|\text{representative state})$ appears similar to the target statistic, its values (100% and 0%) are not explicitly stated in the description of the DGP and therefore not as salient as the other statistics in the information environment.⁵⁰ Analogously, the realized signal (realized stock price growth) is no longer similar to the target statistic $\Pr(\text{revenue up}|\text{realized price})$. Therefore, applying Prediction 2, both Exact Representativeness and Naive Extrapolation should become less prevalent in *Forecast Revision*.

Table E1 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 nontarget statistic becomes less similar to the target statistic, people are less likely to use the nontarget statistic as their answer to the problem. We also find that the fraction of answers that satisfy the no inference-forecast gap condition 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 the inference problems and the forecast-revision problems. The changes in modal behavior also alter the updating bias in the aggregate: Table E2 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.⁵¹

⁵⁰Specifically, participants are told that “Good firms’ revenues grow every month. Bad firms’ revenues never grow in any month.”

⁵¹One may notice that in both *Less Similar* and *Deterministic Outcome* discussed in Section 3.2, the signal and the target statistic in *Forecast Revision* and are framed as two different variables (i.e., stock price and revenue). However,

Table E1: Modes of behavior in *Less Similar*

Mode	<i>Inference</i>	<i>Forecast Revision</i>
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.

Table E2: Aggregate patterns in *Less Similar*

<i>N</i> =60, <i>Obs</i> =467	Classification			Update
	Underreaction	Near-rational	Overreaction	Mean (s.e.)
<i>Inference</i>	64.7%	12.2%	23.1%	14.3 (1.6)
<i>Forecast Revision</i>	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.

While the evidence from the *Less Similar* treatment is consistent with our model, alternative explanations exist. One concerns representational complexity (Ba, Bohren, and Imas, 2025): because the outcome variable is now binary rather than continuous as in the baseline experiment, this reduces representational complexity and induces greater underreaction in belief formation.

forecast revisions underreact in *Less Similar*, but overreact in *Deterministic Outcome*. These different results can also be reconciled by our framework. Unlike in *Less Similar*, the nontarget statistic $\mathbb{E}[\text{outcome}|\text{representative state}]$ remains a salient cue in *Deterministic Outcome*, and it is still similar to the target statistic in *Forecast Revision*— $\mathbb{E}[\text{outcome}|\text{realized signal}]$. As a result, Exact Representativeness remains a prevalent heuristic in *Deterministic Outcome*.

F 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 (Federal Reserve Bank of Philadelphia, n.d.). 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-update and Naive Extrapolation. Let y_t denote the year-over-year growth rate of real GDP in quarter t and 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

$$w_{i,t} := \frac{f_{i,t}y_{t+1} - f_{i,t-1}y_{t+1}}{y_t - f_{i,t-1}y_{t+1}}, \quad (14)$$

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 F1 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-update. 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; UBS/Gallup, n.d.). 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 (Center for Research in Security Prices (CRSP), n.d.) of the most recent year ($r_{t-12,t}$) and the prior forecast before that, we can identify the heuristics of Non-update 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 $t - 60$ to $t - 12$

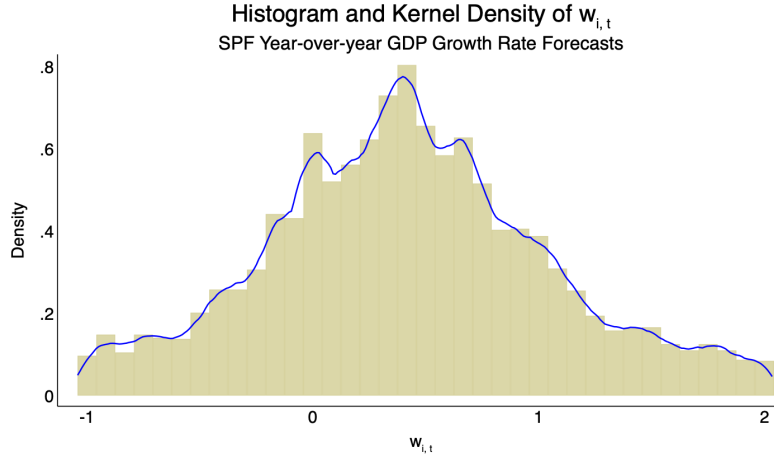


Figure F1: Histogram and kernel density for the weight on the most recent realization (relative to the prior forecast) in the *Survey of Professional Forecasters* (SPF) annual GDP growth rate forecasts, from 1968:Q4 to 2022:Q4. The width of both the bars and the kernel is $\frac{1}{12}$.

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

$$w_{i,t} := \frac{f_{i,t} r_{t,t+12} - r_{t-60,t-12}}{r_{t-12,t} - r_{t-60,t-12}}, \quad (15)$$

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-update and Naive Extrapolation are identified by this measure being close to 0 and 1, respectively.

Figure F2 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-update and Naive Extrapolation are important drivers of stock market return expectations.

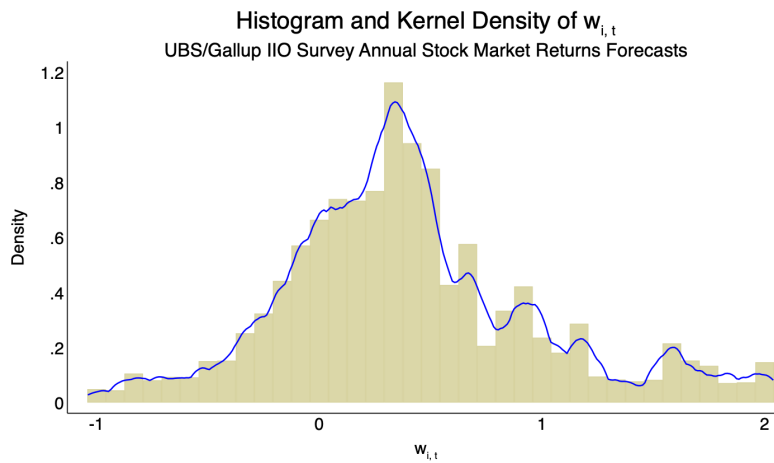


Figure F2: Histogram and kernel density for the weight on the most recent realization (relative to the prior forecast) in annual stock market returns forecasts in the UBS/Gallup *Index of Investor Optimism* (IIO) survey from 1998 to 2007. The width of both the bars and the kernel is $\frac{1}{12}$.