The Inference-Forecast Gap in Belief Updating*

Tony Q. Fan[†]
(Job Market Paper)

Yucheng Liang[‡]

Cameron Peng[§]

[Link to latest version]

November 19, 2022

Abstract

Individual forecasts of economic variables show widespread overreaction to recent news, but laboratory experiments on belief updating typically find underinference from new signals. We provide new experimental evidence to connect these two seemingly inconsistent phenomena. Building on a classic experimental paradigm, we study how people make inferences *and* revise forecasts in the same information environment. Participants underreact to signals when inferring about underlying states, but overreact to signals when revising forecasts about future outcomes. This gap in belief updating is largely driven by the use of different simplifying heuristics for the two tasks. Additional treatments suggest that the choice of heuristics is affected by the similarity between cues in the information environment and the belief updating question: when forming a posterior belief, participants are more likely to rely on cues that appear similar to the variable elicited by the question.

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

[†] Stanford University.

[‡] Carnegie Mellon University.

[§] London School of Economics and Political Science.

1 Introduction

When new information arrives, rational agents should update their beliefs according to Bayes' rule. Empirical research, however, has uncovered many instances in which agents' reactions to information deviate from Bayes' rule. One recurring theme in the existing literature is that the type of belief-updating biases appears to vary from setting to setting. For instance, excess volatility in financial markets and boom-bust cycles in the macroeconomy are more consistent with overreaction to information (e.g., Barberis et al., 2015; Maxted, 2020; Bordalo et al., 2021b). In contrast, post-earnings announcement drifts and households' sluggish responses to macroeconomic conditions can be better understood with underreaction to information (e.g., Barberis et al., 1998; Coibion and Gorodnichenko, 2015). This observation is further echoed in research that directly elicits beliefs and belief changes in both laboratory and field settings: while some studies find clear evidence of underreaction, others find the opposite pattern (see a more detailed review below).

Both overreaction and underreaction are useful concepts in economic analysis and have spurred the development of theories tackling important puzzles in finance and macroeconomics. However, so far we still know little about what makes people overreact in some cases but underreact in others (Benjamin, 2019). Answering this question requires uncovering factors that moderate the direction and magnitude of belief-updating biases. Progress on this front can shed light on the cognitive foundations of information processing and add more discipline and predictive power to models that assume non-Bayesian updating.

In this paper, we propose one condition that mediates underreaction and overreaction to new information. It is motivated by an apparent tension between two large literatures that directly test Bayesian updating using reported beliefs. On the one hand, in both field and laboratory settings, individuals often overreact to recent news when asked to make forecasts (e.g., Hey, 1994; Greenwood and Shleifer, 2014; Gennaioli et al., 2016; Frydman and Nave, 2017; Conlon et al., 2018; Afrouzi et al., 2020; Bordalo et al., 2020). On the other hand, when asked to make inferences about underlying states, participants in experiments typically underreact to realized signals (see Benjamin (2019) for a detailed review). While this tension may be attributed to differences in con-

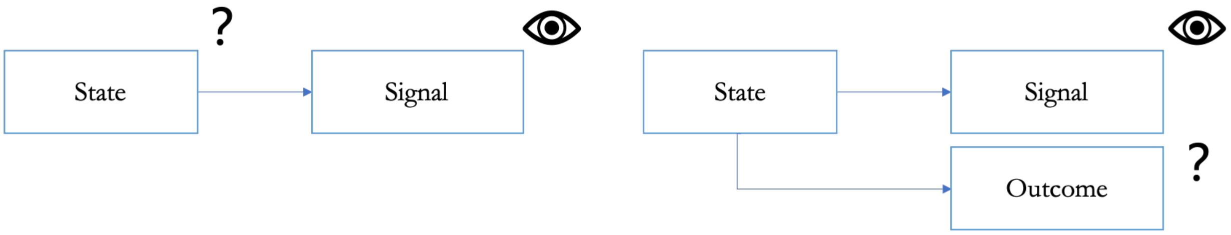


Figure 1: Inference problem (left) and forecast-revision problem (right)

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.

texts or data-generating processes (DGPs), we propose an alternative explanation that has previously been neglected: belief updating differs between an inference problem and a forecast-revision problem. The differences between the two problems are illustrated in Figure 1. An inference problem is one where an agent observes signals and learns about *the underlying state* that determines the distribution of signals. By contrast, a forecast-revision problem is one where an agent also observes signals but instead update beliefs about *future outcomes* whose distributions depend on the underlying state.

In standard models, the forecast-revision problem is closely tied to the inference problem: inference about the underlying state is often the first step or input to revising forecasts about future outcomes. However, we uncover a disconnect between the two: by conducting a series of controlled experiments in which participants perform both types of updating tasks, we show participants underreact to signals when making inferences but overreact when revising forecasts. This finding can reconcile the seemingly inconsistent stylized facts in the aforementioned empirical literature.

Our baseline treatment follows the "bookbag-and-poker-chip" paradigm[1] in experimental research but frames the relevant variables in economic terms. In each round of the experiment, there is a "firm" with a fixed state which is either good or bad. The firm generates signals, framed as its

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

monthly stock price growth, and the signals are informative of the state; good firms, on average, have a higher growth in stock price than bad firms. Participants do not know the true state but are given the full DGP, including the prior distribution over the two states and the distributions of signals conditional on each state. In each month, the signal distribution is i.i.d. normal, with a mean of 100 if the state is good and 0 if it is bad.

The key to our experimental 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 then 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 tightly linked: if one believes that the firm is good with a $p\%$ chance, then by the Law of Iterated Expectations (LoIE), the expectation about the next signal should be $p\% \times 100 + (1 - p\%) \times 0 = p$. The simplicity of this relationship ensures that, for participants who understand this link, the two problems involve a similar level of computational complexity.

Despite the straightforward connection between *Inference* and *Forecast Revision*, participants' behaviors exhibit distinct patterns in the two tasks. In *Inference*, 61% of the answers underreact relative to the Bayesian benchmark while 24% overreact, a result that replicates the stylized fact of systematic underreaction in the bookbag-and-poker-chip literature. By contrast, in *Forecast Revision*, 40% of the answers underreact while 53% overreact. Similarly, when belief updates are measured using the difference between posterior and prior beliefs, the average magnitude of belief updates is substantially larger for *Forecast Revision* than for *Inference*. We refer to this discrepancy in belief updating as the “inference-forecast gap.” This gap is robust across subsamples, across rounds, and under alternative framings of the signal and the outcome. Moreover, the gap persists in two additional treatments: one in which the signal follows a binary distribution and one in which the outcome is different from the signal and completely determined by the state. These treatments not only demonstrate that the gap is robust to alternative DGPs, but also help rule out explanations based on, for example, misperceptions of signal autocorrelation and related phenomena such as the hot-hand bias.

After documenting the inference-forecast gap, we further examine the decision procedures used by participants in the experiment. 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 then (b) using these posterior beliefs to compute the expected value of the forecast outcome under the LoIE. The existence of the gap suggests the use of alternative, nonstandard decision procedures in *Forecast Revision*. One possibility is that participants intend to follow the infer-then-LoIE procedure, but make errors or take shortcuts due to its complexity. 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. Taken together, these results suggest that participants do not appear to be using the infer-then-LoIE procedure when solving forecast-revision problems—correctly or with errors. Instead, they resort to alternative procedures.

What alternative decision procedures do participants use? We shed light on this question by detecting potential modal behaviors in the two updating tasks. In *Inference*, the modal behavior is "non-updates:" in 30% of the answers, the posterior equals the prior. In *Forecast Revision*, the fraction of non-updates drops to 23%; meanwhile, two other behaviors that rarely appear in *Inference* become modal. Under the first mode, which represents 21% of the answers, participants answer 100 when the signal is good and 0 when it is bad. These participants make forecasts as if they were 100% sure about being in the more representative state—that is, 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. That is, participants directly use the realized signal as their expectation of the next outcome—a simplifying heuristic we term “naive extrapolation.” Each of the three modal behaviors corresponds to participants using a different salient cue in the information environment—the prior, the outcome expectation conditional on the representative state, and the realized signal—as an anchor in making forecasts (Kahneman and Frederick, 2002; Shah and Oppenheimer, 2008). These modal behaviors are also important drivers of the aggregate result; excluding them would largely reduce the inference-forecast gap.

Why do participants use different simplifying heuristics, even when the information environment remains unchanged? Building on the literature on salience and memory retrieval (Gennaioli and Shleifer, 2010; Kahana, 2012; Bordalo et al., 2021a), we hypothesize that when answering a belief updating question, people are more likely to rely on salient cues that appear similar to the variable elicited by that question. For example, the expected outcome conditional on the representative state is a salient cue in the information environment. Moreover, this variable appears more similar to the expected outcome conditional on the signal (the forecast-revision variable) than to the posterior probabilities of the states (the inference variable). Therefore, our similarity-based hypothesis predicts that participants are more likely to anchor on this cue when they revise forecasts, which explains the prevalence of exact representativeness. Analogously, the realized signal as a cue is more similar to the forecast-revision variable than to the inference variable, so it is more likely to serve as an anchor when participants revise forecasts, resulting in the behavioral mode of naive extrapolation.

To further test this similarity-based hypothesis, we run two additional treatments. In the first treatment, we reframe the information environment and the belief updating questions in order to *increase* the similarity between the inference variable and the two cues driving exact representativeness and naive extrapolation. This change makes these two heuristics more prevalent among inference answers and reduces the inference-forecast gap. In the second treatment, we reframe the forecast-revision question to *decrease* the similarity between the forecast-revision variable and the two cues. Consistent with our hypothesis, exact representativeness and naive extrapolation become less prevalent among forecast-revision answers, and the inference-forecast gap disappears.

Overall, these treatments support the view that the similarity between cues in the information environment and the variable elicited in the belief-updating question can offer a unifying explanation for underreaction and overreaction as well as the heuristics that drive them.

Our work is related to an active body of experimental research on the conditions of overreaction and underreaction in belief updating (Afrouzi et al., 2020; Enke and Graeber, 2020; He and Kucinskas, 2020; Enke et al., 2021; Hartzmark et al., 2021; Liang, 2021).[2] We replicate the finding from the bookbag-and-poker-chip paradigm that people underreact to information when updating beliefs about underlying states (Phillips and Edwards, 1966; Benjamin, 2019). Importantly, we show that underreaction does not generalize to forecast-revision problems that ask participants to predict future outcomes, even though the information environment does not change.[3] We thus bring a new perspective to this literature; namely, that the direction of belief-updating biases depends on the type of belief elicited. The documented inference-forecast gap is largely due to the use of different simplifying heuristics in the two types of problems. This finding contributes to the vast literature on heuristic decision making (e.g., Tversky and Kahneman, 1974; Shah and Oppenheimer, 2008) and is also consistent with recent evidence on the roles of complexity and incorrect mental models in explaining belief-updating biases (Enke and Zimmermann, 2019; Enke, 2020; Esponda et al., 2020; Andre et al., 2021; Graeber, 2021). Moreover, we build on recent work on salience and memory retrieval (Gennaioli and Shleifer, 2010; Kahana, 2012; Bordalo et al., 2021a) and argue that the similarity between belief-updating questions and salient cues in the information environment plays an important role in reconciling the differential updating behaviors in inference and forecast-revision tasks.[4]

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

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

[4] Our paper is also related to the psychology literature on the asymmetry between diagnostic reasoning ($\Pr(\text{Cause}|\text{Effect})$) and predictive reasoning ($\Pr(\text{Effect}|\text{Cause})$) in a given causal structure (e.g., Tversky and Kahneman, 1980; Fernbach et al., 2011). While the inference problem in our paper is synonymous to diagnostic reasoning, forecast revision is different from either kinds of reasoning in this literature because it elicits the belief of one "effect" (the forecast outcome) of the "cause" (the underlying state) conditional on another effect (the signal). Moreover, in

The finding of overreaction in forecast revisions provides experimental support for overreaction in survey expectations.[5] In this regard, our paper complements studies that find overextrapolation in autocorrelated time-series forecasts (Hey, 1994; Frydman and Nave, 2017; Afrouzi et al., 2020; He and Kucinskas, 2020).[6] DGPs in our experiment, unlike those in these previous studies, fully specify the underlying states, which in turn determine the signal and outcome distributions. This design brings the setting closer to standard models in macroeconomics and finance and lends several advantages to our analysis.[7] First, the explicit separation between states and outcomes makes it possible to design different problems targeting inference and forecast revision, respectively, thereby allowing us to pin down where a specific updating bias arises. Second, it allows us to separately identify the specific forms of overreaction, such as representativeness-based overreaction (Kahneman and Tversky, 1972; Bordalo et al., 2018) and mechanical extrapolation (Barberis et al., 2015, 2018). Third, having a fully-specified DGP allows us to attribute biases in posterior beliefs to incorrect statistical reasoning rather than to misperceived DGPs.

Overreaction in *Forecast Revision* is reminiscent of the hot-hand bias, the exaggeration of belief in an outcome after observing a long streak of the same outcomes (Gilovich et al., 1985; Tversky and Gilovich, 1989; Suetens et al., 2016). In contrast, overreaction occurs in our experiment after just *one* signal realization. Moreover, we find overreaction even when the forecast outcome is different from the signal variable and fully determined by the state, a setting in which misperceptions of outcome autocorrelation and related phenomena such as the hot-hand bias, are irrelevant. Our underinference result is also inconsistent with the leading account of the hot-hand bias, which is based on overinference (Rabin, 2002; Rabin and Vayanos, 2010). At the design level, we use ex-

parts of our experiments, we elicit forecasts without showing participants any signal, which is more akin to predictive reasoning. However, we show that biases in these parts cannot explain the inference-forecast gap. We thank Thomas Graeber for pointing us to this literature.

[5] For example, see Greenwood and Shleifer (2014); Gennaioli et al. (2016); Conlon et al. (2018); Bordalo et al. (2020); Barrero (2022); and Kohlhas and Walther (2021).

[6] In addition to overreaction in autocorrelated time-series forecasts, He and Kucinskas (2020) also finds that forecasts underreact to past observations of a different variable.

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

plicit instructions and comprehension checks to make sure participants do not commit the hot-hand bias. Overall, it is unlikely that our results are driven by the hot-hand bias.

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

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 realized signal.

The main treatment, *Baseline*, has five parts, summarized in Table 1. Each part has eight rounds of problems. In each round, participants are first presented with a "firm" randomly drawn from a new pool of 20 firms. A firm's state, θ, is either G(ood) or B(ad). Participants do not know the state of the drawn firm, but are given the composition of the pool, which specifies the prior distribution over the states. The firm generates signals, s_t, which are framed as the firm's stock price growth in month t. Participants are provided with the conditional distributions of signals: signals of a good firm follow an i.i.d. normal distribution of $N(100, \sigma^2)$ and signals of a bad firm follow i.i.d. $N(0, \sigma^2)$.[8] 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, in each round, the DGP is fully specified by two pieces of information: the prior distribution of states and the conditional distributions of signals. Both are presented to participants

[8] In 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	*Inference Prior*	No	$\Pr(\theta)$
2	*Inference*	Yes	$\Pr(\theta \vert s_0)$
3	*Forecast Prior*	No	$\mathbb{E}(s_1)$
4	*Forecast Revision*	Yes	$\mathbb{E}(s_1 \vert s_0)$
5	*Expectation Formation*	No	$\mathbb{E}(s_1)$

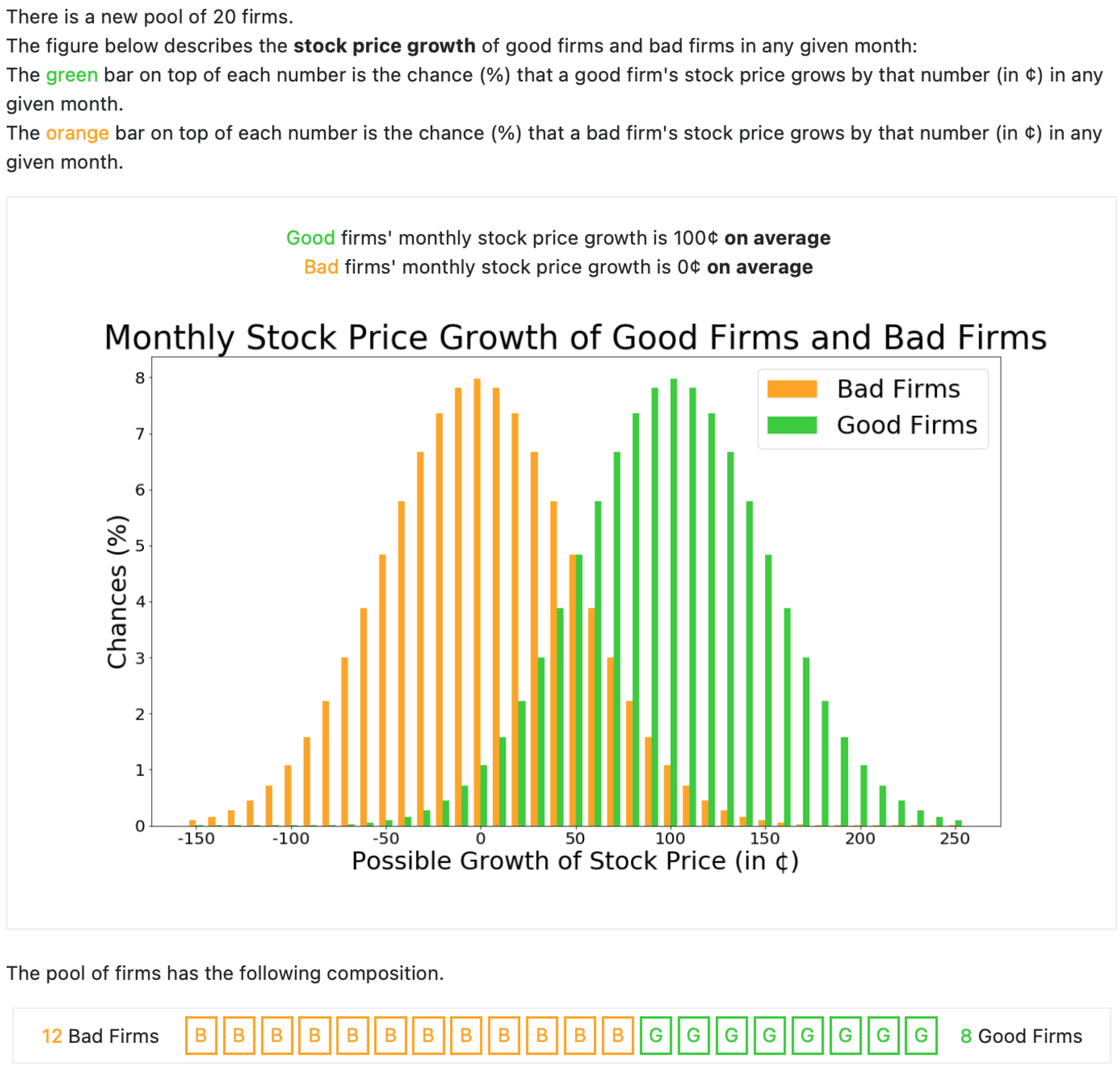
Table 2: Parameter values for DGPs

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

using figures and texts in a one-page display (see Figure 2 for an example), and we explain this interface with detailed instructions.[9] Table 2 summarizes the parameter values for the eight DGPs. We include six DGPs with symmetric priors ($\Pr(G) = 50\%$) and two DGPs with asymmetric priors. The DGPs with symmetric priors allow us to identify underreaction and overreaction without confounds from base-rate neglect, while the DGPs with asymmetric priors help us examine the robustness of our results. Each DGP is represented by one problem in each of the five parts (the DGP is modified in the *Expectation Formation* part, which we will explain later). As a result, answers across parts are directly comparable. Unless mentioned otherwise, an observation refers to a participant's answers to the five corresponding questions in the five parts.

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

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



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.

Good firms' monthly stock price growth is 100¢ **on average**
Bad firms' monthly stock price growth is 0¢ **on average**

Monthly Stock Price Growth of Good Firms and Bad Firms
Bad Firms
Good Firms
Chances (%)
0
1
2
3
4
5
6
7
8
-150
-100
-50
0
50
100
150
200
250
Possible Growth of Stock Price (in ¢)

The pool of firms has the following composition.

12 Bad Firms B B B B B B B B B B B B G G G G G G G G 8 Good Firms

Figure 2: An example of the interface for the DGP

beliefs are elicited in percentages, and henceforth we will refer to an inference answer as the reported belief about the Good state without the $\%$ sign.[10] In *Forecast Revision*, participants instead report their updated expectations about the firm's stock price growth next month $\mathbb{E}(s_1|s_0)$. To allow for a direct comparison between the two parts, the signal realization is set to be the same in any two corresponding rounds for the same participant, though it varies across participants.

In the other three parts, participants do not observe any signal realization before their beliefs are elicited. In *Inference Prior*, participants directly report prior beliefs about the states $\Pr(\theta)$ based on their knowledge about the DGP. Similarly, in *Forecast Prior*, they directly report prior expectations about the signal $\mathbb{E}(s_1)$. These two parts test whether participants can correctly form prior beliefs. The last part, *Expectation Formation*, is identical to *Forecast Prior*, except for the composition of firms in the pool. While the composition of firms in *Forecast Prior* is set exogenously according to Table 2, in *Expectation Formation* it is determined endogenously by participants' reported posterior beliefs about the states in *Inference*. For example, if a participant reports a posterior belief of $\Pr(G|s_0) = 40\%$ in a round in *Inference*, then the pool of firms in the corresponding round in *Expectation Formation* will have 8 ($= 40\% \times 20$) good firms and 12 bad ones.[11] *Expectation Formation* is designed to test whether participants can correctly form expectations about the next signal when the states are distributed according to their own inference posteriors.

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

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

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

After the five parts, participants complete an unincentivized exit survey. At the end of the experiment, participants may receive a \$5 bonus payment, and their chance of receiving the bonus depends on their answer in one randomly selected round through a quadratic rule.[12]

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

2.2 The no inference-forecast gap benchmark

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

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

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

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

We term Equation (2) the "no inference-forecast gap" condition. It summarizes the theoretical link between the posterior belief about the underlying states and the updated forecast of the outcome variable s_1. If an inference answer and its corresponding forecast-revision answer satisfy this con-

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

dition, then there should be no discrepancy between these two types of belief-updating problems: Bayesian inference would translate to rational forecasts, and any deviation from Bayes' rule in the inference answer would imply the same deviation from rationality in the forecast-revision answer.

The computational simplicity of Equation (2) is an advantage of our experimental design. Under the no inference-forecast gap condition, if a signal leads to a belief that the good state has 40% probability, then the resulting expectation of the outcome should be 40. For participants who understand this condition, the computational cost of solving a forecast-revision problem is very close to that of solving the corresponding inference problem. Therefore, computational complexity alone is unlikely to cause violations of the no inference-forecast gap condition.[13]

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

2.3 Instructions and comprehension questions

Participants receive extensive instructions, with the tasks and incentive structure explained in detailed and intuitive terms. In particular, we go to great lengths to ensure that participants fully understand the DGP. First, we emphasize that the state of a firm is constant across months but the signals are i.i.d. conditional on the state. In doing so, we explicitly caution against incorrect beliefs that the signals are autocorrelated conditional on the state. Second, we use an example DGP to illustrate the discretized normal distributions of the signals. In particular, we highlight the conditional means (0 and 100) and the property that signals higher (lower) than 50 are good (bad) news about the firm's quality. Third, we present participants with two explicit formulae, one for

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

calculating the prior distribution over states from the pool composition ($\Pr(G) = \frac{\text{number of Good firms}}{20}$) and one for calculating the expectation about the signal from the belief about the states ($\mathbb{E}(s) = \Pr(G) \times 100$). However, we do not mention or nudge participants toward any specific belief-updating rule.

At the end of the instructions, participants need to answer a set of comprehension questions that test their understanding of the DGP, the incentive structure, and the two formulae. Participants can proceed only if they have answered all the comprehension questions correctly.[14]

2.4 Procedural details

We programmed our experiment using oTree (Chen et al., 2016). For *Baseline*, we recruited 279 participants through Prolific, an online platform designed for social science research.[15] Signals were framed as monthly revenue growth for 142 participants and as stock price growth for 137 participants. There was also some variation across participants in the order of parts: 102 participants went through the experiment in the order of 12345, 103 in the order of 12534, and 74 in the order of 34125. The participants, on average, spent about 30 minutes on the experiment and earned a payment of \$7.08, \$5 of which was the base payment.

2.5 Other treatments

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

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

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

Table 3: Overview of additional treatments

Treatment	Section	Key differences from *Baseline*
Deterministic Outcome	3.2	Forecast outcome is a different variable (100 if $\theta = G$ and 0 if $\theta = B$)
Binary Signal	3.3	Signals are binary; forecast questions ask about full distributions
Nudge	4.1	Beliefs about states and forecasts are elicited on the same page
More Similar	5.1	State variable (profitability) = mean of signal or forecast outcome (profits); inference questions ask about the expectation of the state
Less Similar	5.2	Forecast outcome is a different variable (up if $\theta = G$ and down if $\theta = B$); forecast questions ask about full distributions

3 Evidence for the Inference-Forecast Gap

3.1 Aggregate patterns

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

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

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

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

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

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

We first classify answers in *Inference* and *Forecast Revision* by how they compare to the rational benchmarks. An answer is classified as Near-rational if its difference from the rational benchmark is no more than 2.5.[16] To introduce the categories of Underreaction and Overreaction, we first define an "update" by how much an answer moves from its (objective) prior value in the direction of the realized signal s_0:

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

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

Table 4 shows the aggregate patterns in *Baseline* (excluding observations with a signal of 50). The first three columns concern the distribution of answers by categories. Results from *Inference* replicate the key finding from the classic bookbag-and-poker-chip literature: participants overwhelmingly underreact to new information and update too little about the firm's underlying state. Out of all the answers, 60.5% are Underreaction, 24.2% are Overreaction, and 15.3% are Near-rational. These patterns, however, flip in *Forecast Revision*: 53.3% of the answers indicate

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

Table 4: Aggregate patterns in *Baseline*

N=279, Obs.=2069	Classification			Update
	Underreaction	Near-rational	Overreaction	Mean (s.e.)
Inference	60.5%	15.3%	24.2%	14.6 (.7)
Forecast Revision	40.2%	6.6%	53.3%	31.9 (2)
Rational				23.4 (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.

overreaction to new information, higher than the fraction of 40.2% classified as Underreaction.

The last column of Table 4 concerns the average update. In *Inference*, the average update is 14.6, significantly lower than the average rational update of 23.4 ($p < 0.01$). By contrast, in *Forecast Revision*, the average update is 31.9, significantly higher than the rational benchmark ($p < 0.01$). Therefore, both methods of analysis suggest an inference-forecast gap. In the Appendix, Table A6 further confirms the statistical significance of the inference-forecast gap in a regression framework.

The inference-forecast gap is highly robust in various cuts of the data (see Section A of the Appendix for details). First, in a more "reasonable" subsample that only includes observations with (a) answers within $[0, 100]$ and (b) updates in the correct direction, *Forecast Revision* no longer exhibits overreaction on average, but the inference-forecast gap remains highly significant. Second, the gap is present under all eight DGPs, even though they entail different priors and signal distributions. Third, the gap increases for stronger signals—that is, when the signal deviates more from 50 and therefore becomes more informative—but exists even for the weakest signals. Fourth, our results persist in a subsample that excludes observations with incorrect reported prior beliefs. Fifth, there is no qualitative impacts on the inference-forecast gap (a) when we change the order of experimental parts, (b) when the signal and outcome are framed as revenue growth, and (c) when we control for participant characteristics. One framing variation that has a significant impact on the magnitude of the gap is the timing of outcome realization. The gap shrinks by over a third when the outcome is framed as stock price of the last month rather than of the next month (see Table A9). This result suggests that beliefs about unrealized outcomes may be more responsive to signals than beliefs about realized outcomes.[17]

3.2 *Deterministic Outcome* treatment

In this and the next subsection, we investigate the inference-forecast gap in two additional treatments with alternative DGPs. Both treatments generate patterns similar to those of *Baseline*. These results demonstrate the prevalence of the inference-forecast gap in various environments and help rule out several potential explanations for its emergence.

In *Baseline*, the forecast outcome and the realized signal are part of the same time series. Therefore, the observed inference-forecast gap could be due to misperceived signal autocorrelation and related phenomena such as the hot-hand bias (Gilovich et al., 1985; Tversky and Gilovich, 1989; Suetens et al., 2016). To rule out this explanation, we implement an additional treatment called *Deterministic Outcome*. In this treatment, the outcome variable in *Forecast Revision* is different from the signal variable: when the outcome variable is the firm's stock price growth, the signal variable is the revenue growth, and vice versa. Moreover, the outcome variable is fully determined by the state: it equals 100 for sure in the Good state and 0 for sure in the Bad state. The distributions of the state and the signal are the same as in *Baseline*. Under this alternative DGP, the no inference-forecast gap condition remains the same: the forecast-revision answer equals the corresponding inference answer (minus the % sign). But unlike in *Baseline*, the perceived correlation between the signal and the outcome should be irrelevant for the inference-forecast gap here: since the outcome is fully determined by the state, the perceived signal-outcome correlation

[17] To the best of our knowledge, our study is the first to uncover the effect of the timing of outcome realization on belief-updating biases. 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.

Table 5: Aggregate patterns in *Deterministic Outcome*

N=100, Obs.=748	Classification			Update
	Underreaction	Near-rational	Overreaction	Mean (s.e.)
Inference	63.8%	15.1%	21.1%	13.8 (1.3)
Forecast Revision	40.6%	8.7%	50.7%	32.9 (3.3)
Rational				23.3 (.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.

should be the same as the perceived signal-state correlation.

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

Results from *Deterministic Outcome* clearly show that the hot-hand bias cannot account for the inference-forecast gap. This further differentiates our results from overreaction in univariate forecasts (Hey, 1994; Frydman and Nave, 2017; Afrouzi et al., 2020) in which exaggerated autocorrelation is a key driving force. Moreover, the treatment helps address two additional robustness issues. First, the inference-forecast gap is not limited to cases where the signal and the outcome share the same variable name and distribution. Second, even when the state variable and the outcome variable share the same distribution, an inference-forecast gap can still arise.

3.3 *Binary Signal* treatment

In a second treatment called *Binary Signal*, the signal s_t follows a binary distribution instead of a continuous distribution. In particular, the signal is framed as the direction of the firm's stock price movement and is either up or down, and the probability of an upward movement is higher if the firm's state is Good. The parameters for the DGPs are listed in Table 6. In the *Forecast Revi-*

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%

sion part of this treatment, the problem asks about the full probability distribution of the outcome $\Pr(s_1)$, instead of the expectation $\mathbb{E}(s_1)$.

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

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

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

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

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

In *Binary Signal*, the three categories—Near-rational, Underreaction, and Overreaction—are defined in the same way as in *Baseline*, except that the categories for forecast-revision answers are

[18] For example, computational complexity could lead to higher degrees of cognitive uncertainty (Enke and Graeber, 2020). This could push forecast-revision answers toward underreaction.

Table 7: Aggregate patterns in *Binary Signal*

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

Notes: The first three columns present the percentages of answers classified as Underreaction, Near-rational, and Overreaction. The last column shows average belief movements in the signal direction from the (objective) priors and their rational benchmark. The updates of forecast-revision answers are normalized by $Pr(\text{up}|G) - Pr(\text{up}|B)$ so that they are comparable to the inference updates. Observations with the signal equal to 50 are excluded. Standard errors are clustered by participant.

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

4 Decision Procedures

4.1 Implementation errors or alternative procedures?

In this section, we examine the decision procedures used by participants in the experiment. As discussed in Section 2.2, the inference-forecast gap should not arise if participants, when answering a forecast-revision question, correctly implement the infer-then-LoIE procedure by: (a) first updating their beliefs about the states, in the same way as in the corresponding inference problem, and (b) then applying the LoIE to form expectations about the forecast outcome. The existence of

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

an inference-forecast gap therefore rejects that participants correctly implement this procedure in *Forecast Revision*.

However, it is possible that participants implement this procedure *incorrectly*: they may intend to follow the infer-then-LoIE procedure, but make errors or take shortcuts because the procedure itself is inherently two-step rather than one-step. For instance, a participant may have limited cognitive bandwidth to do one operation at a time. When solving the one-step inference problem, she may be capable of forming probabilistic beliefs about the states. However, when trying to implement the two-step infer-then-LoIE procedure in solving the forecast-revision problem, her cognitive bandwidth may only allow her to form a binary belief ("the firm is good" or "the firm is bad") in the first step, an error that can lead to overreaction.

If it is indeed the implementation complexity of forecast-revision problems that is driving the inference-forecast gap, then reducing this complexity should reduce the 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 are still on the screen, they are asked to report their expectations about the next signal.[20] For a participant intending to follow the infer-then-LoIE procedure, this design makes a forecast-revision problem no more complex than 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 is quoted in cents, participants can just type in the exact same number.

According to the hypothesis above, the reduction in complexity should mitigate any implementation errors in the procedure and reduce the inference-forecast gap. However, we find that displaying the inference answer 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 before, participants overwhelmingly underreact in *Inference* and on average overreact in *Forecast*

[20] More specifically, participants have to stay on the page for eight seconds before answering each question. The forecast-revision question appears only after the answer to the inference question has been submitted. Participants can revise their answers to the inference question before they submit their answers to the forecast-revision question.

Table 8: Aggregate patterns in *Nudge*

N=99, Obs.=715	Classification			Update
	Underreaction	Near-rational	Overreaction	Mean (s.e.)
Inference	70.6%	10.2%	19.2%	10.3 (1.3)
Forecast Revision	42.2%	6.7%	51.0%	28.9 (2.9)
Expectation Formation	60.6%	6.9%	32.6%	13.7 (2.1)
Rational				22.6 (.5)

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

Revision.[21]

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

Figure 3 plots the average deviation from LoIE in expectation-formation problems by the prior (probability of the Good state) and shows that, on average, the deviation is small in magnitude

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

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

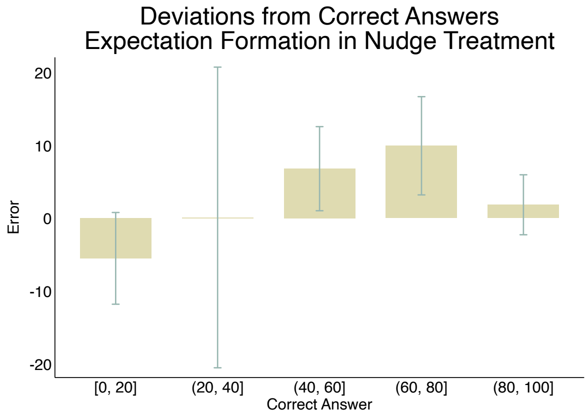


Figure 3: Deviations from LIE in expectation-formation problems

Notes: Standard errors are clustered by participant.

across the board. Moreover, in the third row of Table 8, we classify expectation-formation answers and calculate their updates.[23] Comparing the average update in *Inference*, *Forecast Revision*, and *Expectation Formation*, we find that mistakes in *Expectation Formation* can account for only 18% $(= \frac{13.7-10.3}{28.9-10.3})$ of the inference-forecast gap. Therefore, it is unlikely that the inference-forecast gap stems from the mistakes participants make in standalone expectation-formation problems.

Taken together, results from *Nudge* suggest that the inference-forecast gap does not stem from complexity-induced errors or shortcuts. Therefore, participants do not appear to be following the infer-then-LoIE procedure when solving forecast-revision problems—correctly or with errors. Rather, they appear to be using alternative procedures.

4.2 Alternative decision procedures

What alternative decision procedures do participants use in *Forecast Revision*? To answer this question, we examine the distributions of posterior beliefs to detect potential modal behaviors. To illustrate, Figure 4 plots the answer against the realized signal for problems with symmetric

[23] Similar to before, the update of an expectation-formation answer is defined as the answer minus the (objective) prior in the corresponding forecast-revision problem if the signal in the latter problem is greater than 50 and the reverse if the signal is smaller than 50. The classification of an expectation-formation answer is conducted against the rational benchmark for the corresponding forecast-revision problem.

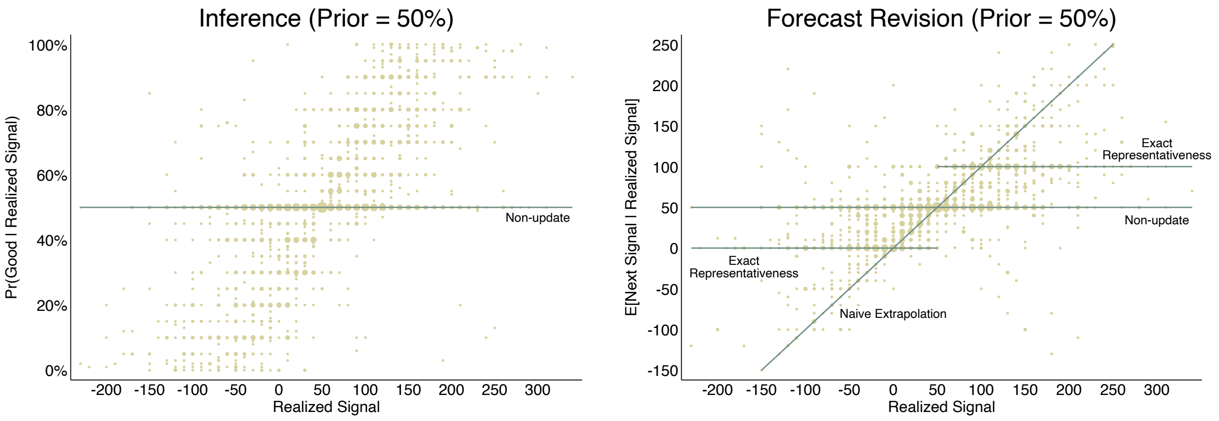


Figure 4: Scatterplots of answers against realized signals: subsample with symmetric priors

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.

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 non-updates replicates a stylized fact in previous inference experiments (e.g., Coutts, 2019; Graeber, 2021).

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

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

Table 9: Modes of behavior in *Baseline*

Mode	Criterion for answer	*Inference*	*Forecast Revision*
Non-update	$=$ prior	29.7%	22.6%
Exact representativeness	$= 100$ if $s_0 > 50$, $= 0$ if $s_0 < 50$	2.6%	20.5%
Naive extrapolation	$= s_0$	3.1%	11.8%
No inference-forecast gap (excluding the other modes)	inference $=$ forecast revision	3.3%	
Unclassified		61.6%	44.3%
Observations		2069	2069

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.

Second, a smaller yet still significant fraction of forecast-revision answers are anchored at the face value of the realized signal.[25] We term this behavior "naive extrapolation" because it suggests a particular form of extrapolative beliefs whereby participants directly (and naively) use the most recent realization as their forecast for the next realization (Barberis et al., 2015, 2018; Liao et al., 2021).[26] This behavior leads to overreaction in the problems with symmetric priors in our experiment.

In Table 9, we define the behavioral modes and quantify their prevalence in *Baseline*. Confirming the patterns in the scatterplots, non-updates are widespread in both *Inference* and *Forecast Revision*, making up 29.7% and 22.6% of all answers, respectively. The other two behavioral modes, exact representativeness and naive extrapolation, appear almost exclusively in *Forecast Revision*, making up 20.5% and 11.8% 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

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

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

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.[27] 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 inference-forecast gap is in fact reversed when the two modes are excluded, suggesting that the gap is largely explained by the presence of these modes. More details are reported in Tables A6 and A7 of the Appendix.

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

5 Mechanisms

The use of simplifying heuristics *per se* is not surprising given the complexity of the belief-updating tasks. The more surprising observation is the use of different heuristics for solving infer-

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

ence and forecast-revision problems, even though the information environment remains the same. Building on the literature on salience and memory retrieval (Gennaioli and Shleifer, 2010; Kahana, 2012; Bordalo et al., 2021a), we hypothesize that the choice of simplifying heuristics is driven by the similarity between salient cues in the information environment and the variable elicited by the belief-updating question. When the similarity increases, participants are more likely to use that salient cue as an anchor to form their posterior beliefs.[28]

This similarity-based mechanism offers a unifying explanation for the different heuristics observed in *Inference* and *Forecast Revision*, as we summarize in Table 10. For example, in *Forecast Revision*, the question asks participants to make forecasts about the stock price growth in the next period conditional on the realized signal. The elicited variable, expected price growth conditional on the realized signal, is similar to expected price growth conditional on the representative state: both are values of the outcome variable and are expectations conditioned on the realized signal in some way. This similarity may induce participants to use expected price growth conditional on the representative state as an anchor in making forecasts, resulting in exact representativeness when participants do not subsequently adjust it. In contrast, in *Inference*, the question asks about the conditional probability distribution over the states, which appears less similar to expected price growth (conditional on the representative state). As a result, exact representativeness is rarely observed.

A similar argument can be made to explain the different prevalence of naive extrapolation in the two types of updating problems. The realized signal and the elicited variable in *Forecast Revision* are both measures of the firm's stock price growth. If participants use the realized signal as an anchor and do not adjust it, it will result in naive extrapolation in forecasts. In contrast, the realized signal is less similar to the conditional probability distribution over the states, which is the *Inference* variable. As a result, we rarely observe naive extrapolation in *Inference* problems. We can also use similarity to explain the prevalence of non-updates in both types of updating problems: prior beliefs over states and prior outcome expectations are both similar to their posterior

[28] The memory literature suggests that similarity is a key force in memory recall. In particular, experiences that share common features with the present cue are more "available" to be recalled and therefore have a greater influence on decisions (Kahana, 2012; Bordalo et al., 2021a). In our setting, cues are that similar to the question could be more likely to enter participants' working memory and therefore affect their beliefs as a salient cue (Afrouzi et al., 2020).

Table 10: Similarity between belief-updating questions and cues in *Baseline*

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

counterparts. If participants use the prior as the anchor and do not adjust it, it will result in non-updates.

This similarity-based mechanism also suggests that when the similarity between cues and the elicited variable changes, the prevalence of different simplifying heuristics will change as well. Below, we design two treatments which manipulate the similarity between cues and elicited variables by varying the framing of variables and questions.

5.1 *More Similar* treatment

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

Table 11 summarizes the similarity between cues and elicited variables in *More Similar*. The key difference from *Baseline* is the increased similarity between $\mathbb{E}(\text{profitability}\mid\text{realized profit})$,

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

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

the variable elicited in the inference question, and two cues: $\mathbb{E}$(profit|representative state), the expected profit conditional on the representative state, and the realized profit. This increase in similarity comes from the fact that the inference variable and the two cues are now all profit-related measures that are conditioned in some way on the realized signal. According to our hypothesis, participants should be more likely to anchor their inference answers on these two cues. Therefore, we predict that exact representativeness and naive extrapolation in *Inference* will now be more prevalent than before.

Results from *More Similar* support this prediction. Table 12 shows that exact representativeness and naive extrapolation become modal behaviors in *Inference* under this treatment. This is in stark contrast with *Baseline* where these two behaviors are almost non-existent in the same part. This treatment also generates a different aggregate pattern from *Baseline* (see Table 13): the numbers of underreacting and overreacting answers are approximately the same in *Inference*, and the average inference update is even on the overreaction side. The inference-forecast gap also becomes smaller.[29]

[29] In the Appendix, Table A10 shows in a regression that the inference-forecast gap is still significant. This suggests that while making the state variable a monetary performance measure and asking about its expectation can increase the responsiveness to signals in inference problems, these framing changes do not account for the entire inference-forecast gap.

Table 12: Modes of behavior in *More Similar*

Mode	*Inference*	*Forecast Revision*
Non-update	33.3%	31.2%
Exact representativeness	17.9%	19.9%
Naive extrapolation	22.6%	31.4%
No inference-forecast Gap (excluding the other modes)	3.8%	
Unclassified	25.8%	17.4%
Observations	442	442

Notes: The criterion for an answer to be classified into a mode is the same as in Table 9. 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
N=60, Obs=442	Underreaction	Near-rational	Overreaction	Mean (s.e.)
Inference	48.6%	4.8%	46.6%	30.1 (4.4)
Forecast Revision	38.0%	2.9%	59.0%	41.6 (5)
Rational				24.4 (.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.

5.2 *Less Similar* treatment

In a second treatment called *Less Similar*, we reframe the forecast-revision question so that the elicited variable appears *less* similar to two salient cues. The state variable, the signal, and the inference question are the same as in *Baseline*. What is different is that in *Forecast Revision*, after observing the realized stock price growth, participants are asked about the probability that the firm's revenue will go up next month. The direction of the firm's revenue movement is fully

Table 14: Similarity between belief-updating questions and cues in *Less Similar*

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

determined by the state—participants are told that a firm's revenue always goes up if the state is Good and it always goes down if the state is Bad.

Table 14 summarizes the similarity between the cues and the elicited variables in *Less Similar*. In this treatment, exact representativeness can arise if participants anchor at one of two cues, both taking the values of 100 or 0: $\mathbb{E}(\text{price}\mid\text{representative state})$, the expected stock price growth conditional on the representative state, and $\Pr(\text{revenue up}\mid\text{representative state})$, the probability of the revenue going up conditional on the representative state. Compared to in *Baseline*, the first cue $\mathbb{E}(\text{price}\mid\text{representative state})$ is less similar to $\Pr(\text{revenue up}\mid\text{realized price})$, the variable elicited by the forecast-revision question, as the latter is now a probability distribution over revenue movements. For the second cue $\Pr(\text{revenue up}\mid\text{representative state})$, although it appears similar to the elicited variable, its values (100% and 0%) are not explicitly stated in the description of the DGP and therefore not as salient as the other cues in the information environment.[30] Therefore, we predict that exact representativeness will become less prevalent. Relatedly, the realized signal (stock price growth in the current month) is no longer similar to the elicited variable (probability of the revenue going up), and we predict that naive extrapolation will also show up less.

Table 15 shows the modal answers in *Less Similar*. Consistent with our prediction, exact representativeness and naive extrapolation are much less prevalent in *Forecast Revision* compared with *Baseline*. This change in modal behaviors supports our hypothesis that when a cue becomes

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

Table 15: Modes of behavior in *Less Similar*

<table>
<tr><th>Mode</th><th><i>Inference</i></th><th><i>Forecast Revision</i></th></tr>
<tr><td>Non-update</td><td>31.7%</td><td>30.8%</td></tr>
<tr><td>Exact representativeness</td><td>9.0%</td><td>13.9%</td></tr>
<tr><td>Naive extrapolation</td><td>3.9%</td><td>3.6%</td></tr>
<tr><td>No inference-forecast Gap
(excluding the other modes)</td><td colspan="2">11.8%</td></tr>
<tr><td>Unclassified</td><td>45.2%</td><td>41.5%</td></tr>
<tr><td>Observations</td><td>467</td><td>467</td></tr>
</table>

Notes: The criterion for an answer to be classified into a mode is the same as in Table 9. The percentages are the fractions of answers in each mode. Observations with the signal equal to 50 are excluded.

less similar to the question, people are less likely to use heuristics that rely on this cue. Another pattern is that the fraction of answers that satisfy the no inference-forecast gap condition increases from 3.6% in *Baseline* to 11.8% in *Less Similar*. One possible explanation for this result is that the design of *Less Similar* makes it easier for some participants to recognize the tight conceptual connection between inference problems and forecast-revision problems. The change in modal behavior also alters the aggregate pattern of the inference-forecast gap. Table 16 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.

One may notice that in both the *Less Similar* treatment and the *Deterministic Outcome* treatment in Section 3.2, the outcome in *Forecast Revision* and the signal are two different variables. However, the forecast-revision answers in these two treatments exhibit very different patterns: while forecasts underreact in *Less Similar*, they overreact in *Deterministic Outcome*. These different results can also be reconciled by our hypothesis. Unlike in *Less Similar*, the expected outcome conditional on the representative state remains a salient cue in *Deterministic Outcome* and it is still similar to the elicited forecast variable (the expected outcome conditional on the realized signal). As a result, exact representativeness remains a prevalent heuristic in the forecast-revision problems

Table 16: Aggregate patterns in *Less Similar*

N=60, Obs=442	Classification			Update
	Underreaction	Near-rational	Overreaction	Mean (s.e.)
Inference	64.7%	12.2%	23.1%	14.3 (1.6)
Forecast Revision	62.1%	12.8%	25.1%	13.6 (1.8)
Rational				23.1 (.6)

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

in *Deterministic Outcome*, which drives the overreaction.

Taken together, the two similarity treatments suggest that the similarity between cues in the information environment and the elicited variables in belief-updating questions can offer a unifying mechanism for underreaction and overreaction as well as the heuristics that facilitate them.

6 Concluding Remarks

In this paper, we show that people react more to new information when revising forecasts about future outcomes than when making inferences about underlying states, even when the information environment remains the same. Through a series of subsample analyses and additional treatments, we show that the gap is robust to order effects, participant characteristics, and alternative data-generating processes. Therefore, it offers a new perspective to the study of belief-updating biases: the type of bias not only depends on the information environment, but also hinges on the belief-updating question itself.

The fact that the inference-forecast gap we document is largely driven by simplifying heuristics begs more examination of the underlying mechanisms. We show, in two treatments, that similarity is key to explaining the different heuristics observed in inference problems and forecast-revision problems. When the similarity between salient cues in the information environment and the variable elicited by the belief-updating question changes, the use of heuristics changes correspondingly. This result further highlights the role played by salience and memory retrieval in belief formation (Gennaioli and Shleifer, 2010; Kahana, 2012; Bordalo et al., 2021a).

Our results have implications for three broad settings in which underreaction and overreaction have been shown to coexist: experiments, survey expectations, and other field settings such as asset return predictability.[31] In the experimental setting, bookbag-and-poker-chip inference experiments often find underreaction while forecast experiments typically find overreaction. Our results can immediately speak to this discrepancy: due to different levels of similarity between salient cues in the information environment and the variable elicited by the belief-updating question, people use different simplifying heuristics when solving inference and forecast-revision problems.

In the settings of survey expectations and markets, our experiment sheds light on why overreaction, rather than underreaction, is more commonly observed among survey forecasters, especially at the forecaster level (Bordalo et al., 2020). It is worth noting that our results are relevant to the setting of survey expectations even when forecasters are professionals. In reality, the DGPs of key macroeconomic and financial variables are much more complex than the DGPs in our experiment. Even though professional forecasters are on average more sophisticated than the participants we study, they may still need to resort to simplifying heuristics as the participants do in our experiment. It is also worth noting that even professional forecasts have subjective inputs (Stark, 2013) and are highly correlated with the expectations of households (Greenwood and Shleifer, 2014) who are also important market participants and closer to the participants in our study.[32] More broadly, our experimental evidence suggests that, in order to explain underreaction and overreaction in the field, it is important to consider the information environment and especially the similarity between salient cues and the variables people form expectations on.

[31] We thank Nick Barberis for raising this point.

[32] Robert Shiller's United States Stock Market Confidence Indices also show that U.S. institutions and individuals exhibit highly correlated beliefs over time.

References

H. Afrouzi, S. Y. Kwon, A. Landier, Y. Ma, and D. Thesmar. Overreaction and working memory. *Working paper*, 2020.

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

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

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

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

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

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

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

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

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

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

P. Bordalo, J. J. Conlon, N. Gennaioli, S. Y. Kwon, and A. Shleifer. Memory and probability. *Working paper*, 2021a.

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

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

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

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

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

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

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

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

B. Enke and T. Graeber. Cognitive uncertainty. *Working paper*, 2020.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Y. Liang. Learning from unknown information sources. *Working paper*, 2021.

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

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

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

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

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

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

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

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

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

L. D. Phillips and W. Edwards. Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, 72(3):346, 1966.

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

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

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

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

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

T. Stark. Spf panelists' forecasting methods: A note on the aggregate results of a november 2009 special survey. *Federal Reserve Bank of Philadelphia*, 2013.

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

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

A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, 185(4157):1124–1131, 1974.

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

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

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

A Robustness of the Inference-Forecast Gap

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

A.1 A more “reasonable” subsample

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

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

N=279, Obs.=1345	Classification			Update
	Underreaction	Near-rational	Overreaction	Mean (s.e.)
Inference	54.3%	17.9%	27.7%	17.8 (.9)
Forecast Revision	42.7%	9.1%	48.2%	24.3 (1.2)
Rational				23.3 (.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, forecast-revision answers that are outside $[0, 100]$, or updates in the wrong direction are excluded. Standard errors are clustered by participant.

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

A.2 Priors and signals

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

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

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

A.3 Order between parts

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

A.4 Participant characteristics

Finally, we examine the heterogeneity of the gap across participant characteristics, such as gender, education, investment experience, familiarity with statistics and economics, and perfor-

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

		Classification			Update
		Underreaction	Near-rational	Overreaction	Mean (s.e.)
$\mathrm{Pr}(G)$ = 50%	*Inference*	71.1%	19.8%	9.2%	18.7 (1.2)
σ = 50	*Forecast Revision*	44.3%	12.1%	43.6%	31.2 (2.4)
(Obs. = 273)	Rational				35.9 (.8)
$\mathrm{Pr}(G)$ = 50%	*Inference*	68.2%	16.7%	15.1%	17.1 (1.2)
σ = 60	*Forecast Revision*	48.4%	6.6%	45%	28.5 (2.8)
(Obs. = 258)	Rational				31.8 (.8)
$\mathrm{Pr}(G)$ = 50%	*Inference*	64.4%	13.6%	22%	15.4 (1.1)
σ = 70	*Forecast Revision*	40.5%	7.2%	52.3%	29 (2.6)
(Obs. = 264)	Rational				26.9 (.8)
$\mathrm{Pr}(G)$ = 50%	*Inference*	64.5%	12.8%	22.6%	13.8 (1.2)
σ = 80	*Forecast Revision*	40.8%	4.5%	54.7%	33.4 (3.2)
(Obs. = 265)	Rational				24.8 (.8)
$\mathrm{Pr}(G)$ = 50%	*Inference*	50%	18.6%	31.4%	16.3 (1.1)
σ = 90	*Forecast Revision*	37.1%	4.2%	58.7%	35.9 (3.2)
(Obs. = 264)	Rational				21.6 (.7)
$\mathrm{Pr}(G)$ = 50%	*Inference*	51.7%	15.6%	32.7%	12.9 (1.2)
σ = 100	*Forecast Revision*	32.3%	8%	59.7%	38.9 (3.3)
(Obs. = 263)	Rational				19.5 (.7)
$\mathrm{Pr}(G)$ = 80%	*Inference*	55.3%	13.1%	31.6%	11.4 (1.5)
σ = 100	*Forecast Revision*	39.8%	3.3%	57%	30.1 (4)
(Obs. = 244)	Rational				12.6 (.6)
$\mathrm{Pr}(G)$ = 20%	*Inference*	58%	11.3%	30.7%	10.1 (1.6)
σ = 100	*Forecast Revision*	37.8%	6.3%	55.9%	27.9 (3.6)
(Obs. = 238)	Rational				12.2 (.6)

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

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

Signal Strength		Classification			Update
		Underreaction	Near-rational	Overreaction	Mean (s.e.)
Weakest	*Inference*	47.7%	23%	29.3%	4.5 (.8)
(Obs. = 239)	*Forecast Revision*	48.1%	11.7%	40.2%	10.3 (1.5)
	Rational				6.5 (.2)
Weak	*Inference*	59.2%	13.6%	27.2%	9.7 (1)
(Obs. = 309)	*Forecast Revision*	44.3%	5.2%	50.5%	19.1 (2.1)
	Rational				15.9 (.2)
Medium	*Inference*	63.8%	10.4%	25.8%	15.1 (1.1)
(Obs. = 279)	*Forecast Revision*	37.6%	5%	57.3%	33.3 (2.7)
	Rational				25.1 (.1)
Strong	*Inference*	64.9%	12.2%	23%	20.4 (1.4)
(Obs. = 296)	*Forecast Revision*	35.1%	4.1%	60.8%	48.8 (4.1)
	Rational				34.4 (.2)
Strongest	*Inference*	64%	25.3%	10.7%	25.6 (1.4)
(Obs. = 356)	*Forecast Revision*	43.3%	11.2%	45.5%	39.8 (3.5)
	Rational				44.7 (.2)

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

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

N=279, Obs.=1496	Classification			Update
	Underreaction	Near-rational	Overreaction	Mean (s.e.)
Inference	57.8%	17.9%	24.3%	15.6 (.8)
Forecast Revision	43.8%	7.7%	48.5%	27.4 (2.2)
Rational				24 (.3)

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

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

		Classification			Update
		Underreaction	Near-rational	Overreaction	Mean (s.e.)
Order: 12345	*Inference*	55.3%	17.7%	27%	15.8 (1.1)
(N = 102)	*Forecast Revision*	37.6%	7.3%	55%	34 (2.8)
(Obs. = 763)	Rational				22.9 (.4)
Order: 12534	*Inference*	59.9%	15.7%	24.3%	14.7 (1.1)
(N = 103)	*Forecast Revision*	40.9%	5.4%	53.7%	32.4 (3.5)
(Obs. = 756)	Rational				23.3 (.5)
Order: 34125	*Inference*	68.5%	11.3%	20.2%	12.6 (1.5)
(N = 74)	*Forecast Revision*	42.7%	7.1%	50.2%	28.4 (4.3)
(Obs. = 550)	Rational				24.3 (.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.

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

A.5 Framing

In different versions of the *Baseline* treatment, we show that the gap is robust to several changes in the framing of the signal and forecast outcome. First, we frame the signal as the firm's revenue growth (rather than stock price growth); we find a quantitatively smaller but still significant gap with this alternative framing. Second, in the three forecast parts, we ask participants to make predictions about the *previous* signal instead of the next signal; we find an inference-forecast gap that is quantitatively smaller but still significant at the 5% level. Table A9 show these results in regressions.

A.6 Regression analyses

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

	Update		
	Full sample	“Reasonable” updates	Correct priors
	(1)	(2)	(3)
Forecast Revision	17.364***	6.512***	11.751***
	(2.198)	(1.219)	(2.448)
Rational Update	1.040***	0.587***	0.951***
	(0.064)	(0.038)	(0.066)
Problem FE	Yes	Yes	Yes
Subject FE	Yes	Yes	Yes
Observations	4138	2690	2992
R^2	0.333	0.469	0.357

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

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

	Update	
	Full sample & excluding two modes	"Reasonable" updates & excluding two modes
	(1)	(2)
Forecast Revision	11.049***	-2.749**
	(2.844)	(1.123)
Rational Update	1.002***	0.440***
	(0.088)	(0.049)
Problem FE	Yes	Yes
Subject FE	Yes	Yes
Observations	2738	1632
R^2	0.342	0.503

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

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

	Update
Forecast Revision	29.612***
	(3.715)
Male × *Forecast Revision*	-4.993
	(4.292)
College × *Forecast Revision*	-2.075
	(4.279)
Investor × *Forecast Revision*	-3.965
	(4.340)
Familiar with Stats × *Forecast Revision*	-6.406
	(4.775)
Familiar with Econ × *Forecast Revision*	-4.765
	(5.115)
High Comprehension × *Forecast Revision*	-8.614**
	(3.788)
Male	0.089
	(1.379)
College	-1.325
	(1.455)
Investor	5.014***
	(1.595)
Familiar with Stats	2.677*
	(1.553)
Familiar with Econ	-1.473
	(1.678)
High Comprehension	4.409***
	(1.509)
Rational Update	1.016***
	(0.063)
Problem FE	Yes
Observations	4138
R^2	0.162

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

Table A9: Heterogeneity of the inference-forecast gap across alternative framing

	Update	
	Stock price vs. revenue	Next vs. last signal
	(1)	(2)
Stock Price × *Forecast Revision*	19.798***	
	(2.857)	
Revenue × *Forecast Revision*	14.984***	
	(3.120)	
Revenue	1.729	
	(1.410)	
Next × *Forecast Revision*		17.364***
		(2.120)
Last × *Forecast Revision*		14.305***
		(1.977)
Last		0.227
		(1.112)
Rational Update	1.021***	0.979***
	(0.064)	(0.047)
Problem FE	Yes	Yes
Observations	4138	7392
R^2	0.147	0.141

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

Table A10: The inference-forecast gap across different treatments

	Update
Baseline × *Forecast Revision*	17.364***
	(2.121)
Deterministic Outcome × *Forecast Revision*	19.198***
	(3.304)
Nudge × *Forecast Revision*	18.640***
	(2.962)
More Similar × *Forecast Revision*	11.559***
	(3.526)
Less Similar × *Forecast Revision*	-0.665
	(1.642)
Deterministic Outcome	-1.167
	(1.555)
Nudge	-4.218***
	(1.552)
More Similar	14.143***
	(4.109)
Less Similar	-0.564
	(1.783)
Rational Update	0.942***
	(0.051)
Problem FE	Yes
Observations	8882
R^2	0.161

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

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

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

Notes: *, **, and *** indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively. Standard errors are clustered by participnt. 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 Problems with asymmetric priors

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

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

Mode	Criterion for answer	*Inference*	*Forecast Revision*
Non-update	= prior	31.5%	20.1%
Exact Representativeness	$= 100$ if $s_0 > 50$, $= 0$ if $s_0 < 50$	2.9%	16.2%
Naive Extrapolation	$= s_0$	3.1%	9.5%
No Inference-Forecast Gap (excluding the other modes)	inference = forecast revision	2.5%	
Unclassified		61.2%	53.3%
Observations		482	482

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

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

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

B.2 Relaxing criteria for classification

Table B2 shows the prevalence of behavioral modes when we relax the classification criteria to allow for errors within $[-4, 4]$. Compared to the results with strict classification criteria (Table 9), the fraction of answers in each mode increases only slightly, and the overall qualitative pattern remains the same.

B.3 Participant–part–level classification

To study the consistency of behavior within each participant, we conduct a classification exercise at the participant-part level. Specifically, a participant is classified into a type in a part (*Inference* or *Forecast Revision*) if more than half of her answers in that part are classified into the corresponding mode. Table B3 shows the joint distribution of types across the two parts. The numbers of participants classified in the two parts are 74 and 112, 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*.

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

Mode	Criterion for answer	*Inference*	*Forecast Revision*
Non-update	$\approx$ prior	32%	23.5%
Exact Representativeness	$\approx$ 100 if $s_0 > 50$, $\approx$ 0 if $s_0 < 50$	6%	21.3%
Naive Extrapolation	$\approx s_0$	3.8%	12.1%
No Inference-Forecast Gap (excluding the other modes)	inference $\approx$ forecast revision	3.9%	
Unclassified		54.8%	41.9%
Observations		2069	2069

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.

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

Inference type *Forecast Revision* type	Non-update	Exact Representativeness	Naive Extrapolation	No Inference-Forecast Gap	Unclassified	Total
Non-update	24	1	1	0	24	50
Exact Representativeness	3	2	0	0	33	38
Naive Extrapolation	9	0	0	0	13	22
No Inference-Forecast Gap	0	0	0	2	0	2
Unclassified	31	0	1	0	136	168
Total	67	3	2	2	206	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.

B.4 Modes of behavior in other treatments

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

Table B4: Modes of behavior in *Deterministic Outcome*

Mode	Criterion for answer	*Inference*	*Forecast Revision*
Non-update	$=$ prior	35.7%	23.3%
Exact Representativeness	$= 100$ if $s_0 > 50$, $= 0$ if $s_0 < 50$	5.3%	20.6%
Naive Extrapolation	$= s_0$	3.9%	13.5%
No Inference-Forecast Gap (excluding the other modes)	inference = forecast revision	4.8%	
Unclassified		51.3%	41%
Observations		748	748

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.

while exact representativeness and naive extrapolation are only prevalent in the latter.

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

Table B6 presents the distribution of modal behaviors in *Nudge*. The fraction of non-updates in *Inference* is 53.4%, a notable increase from the 29.7% in *Baseline*. However, the fraction of non-updates in *Forecast Revision* remains almost the same as in *Baseline*, as does the fraction of answers classified as exact representativeness and naive extrapolation. In addition, the fraction of answers that satisfy the no inference-forecast gap condition increases to 9.2% 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 B5: Modes of behavior in *Binary Signal*

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

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

Table B6: Modes of behavior in *Nudge*

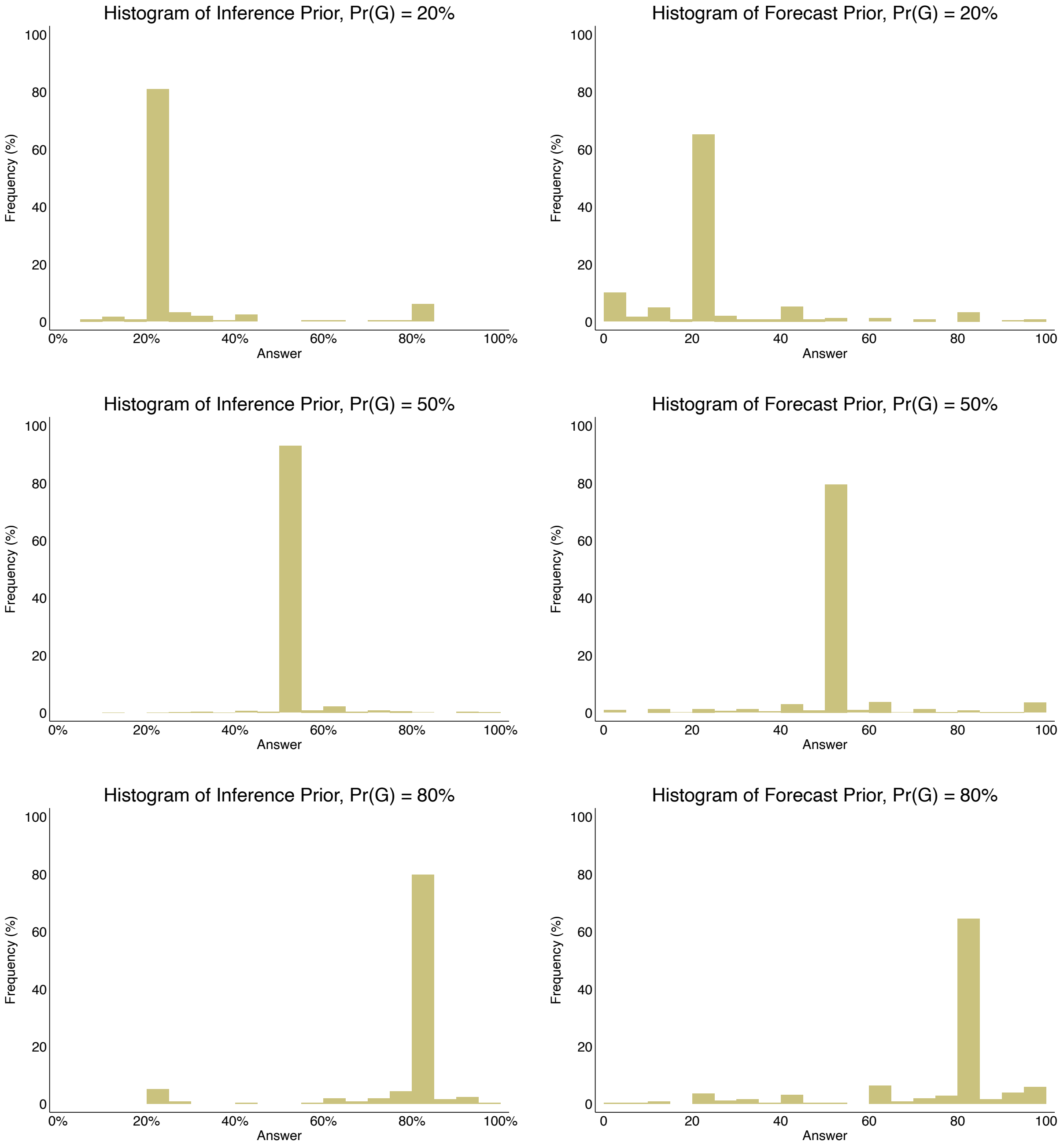
Mode	Criterion for answer	*Inference*	*Forecast Revision*
Non-update	$=$ prior	53.4%	22%
Exact Representativeness	$= 100$ if $s_0 > 50$, $= 0$ if $s_0 < 50$	2.7%	17.9%
Naive Extrapolation	$= s_0$	3.6%	9.1%
No Inference-Forecast Gap (excluding the other modes)	inference = forecast revision	9.2%	
Unclassified		32.6%	44.2%
Observations		715	715

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

C Beliefs without realized signal

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

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



Histogram of Inference Prior, Pr(G) = 20%

Histogram of Forecast Prior, Pr(G) = 20%

Histogram of Inference Prior, Pr(G) = 50%

Histogram of Forecast Prior, Pr(G) = 50%

Histogram of Inference Prior, Pr(G) = 80%

Histogram of Forecast Prior, Pr(G) = 80%

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

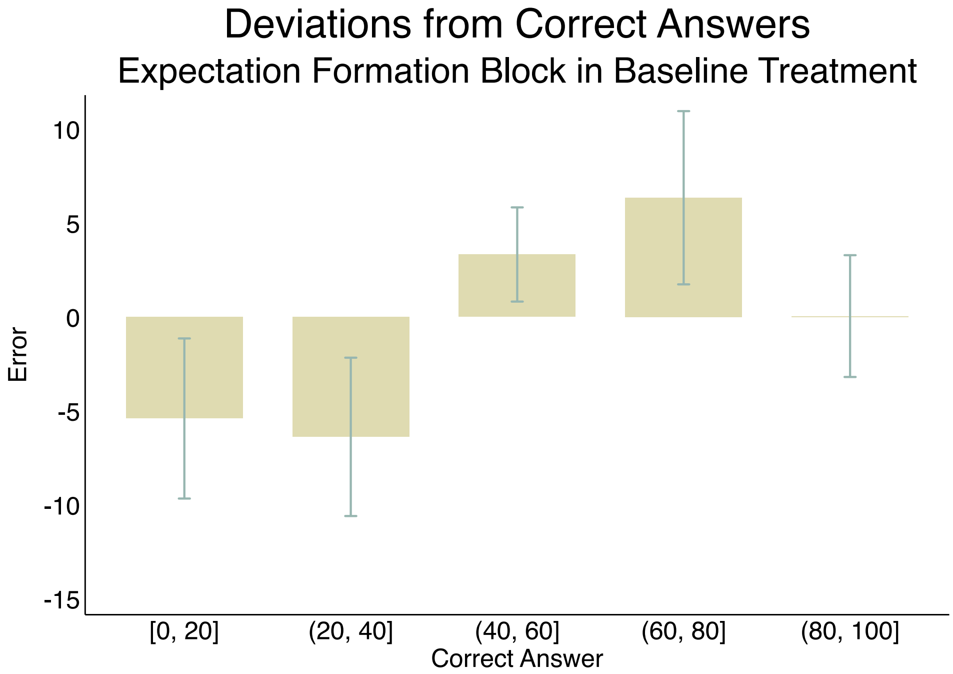


Figure C2: Deviations from LoIE in expectation-formation problems in *Baseline*

Notes: Standard errors are clustered by participant.