The Inference-Forecast Gap in Belief Updating*

Tony Q. Fan† Yucheng Liang‡ Cameron Peng§

March 27, 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, Chris Roth, Josh 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, the current discussion is not satisfying because so far we still know little about what make 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 moderates 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; Bordalo et al., 2020; Afrouzi 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 due to differences in contexts

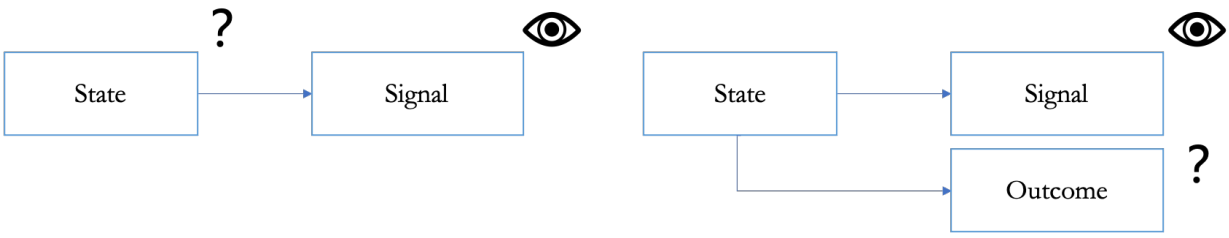

Figure 1: Inference problem (left panel) and forecast-revision problem (right panel)

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.

or data-generating processes (DGPs), we propose an alternative explanation that has previously been neglected: belief updating differs between an inference process and a forecast-revision process. The differences between the two processes are illustrated in Figure 1. An inference process is one where an agent observes signals and learns about an underlying state that determines the distribution of signals. By contrast, a forecast-revision process 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 process closely follows the inference process. However, by conducting a series of controlled experiments in which participants perform both types of updating tasks, we uncover a disconnect between the two: 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 phrases 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 monthly stock price growth, which 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

[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.

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.

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 (LIE), the expectation about the next signal should be $p\% \times 100 + (1 - p\%) \times 0 = p$. The simplicity of this relation ensures that, for participants who understand this link, the two problems pose a similar computational complexity.

Despite the straightforward connection between *Inference* and *Forecast Revision*, participants' behaviors exhibit distinct patterns in the two tasks. In *Inference*, 60% of the answers underreact relative to the Bayesian benchmark while 25% overreact, a result that replicates the stylized fact of systematic underreaction in the bookbag-and-poker-chip literature. By contrast, in *Forecast Revision*, 43% of the answers underreact while 50% 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 such as the hot-hand bias.

After documenting the existence of the inference-forecast gap, we further examine the deci-

sion procedures used by participants in the experiment. The gap should not arise if, in *Forecast Revision*, participants correctly implement the *infer-then-LIE* 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 LIE. The rejection of this standard procedure suggests the use of alternative, nonstandard decision procedures in *Forecast Revision*. One possibility is that participants intend to follow the infer-then-LIE procedure, but make errors or take shortcuts due to the complexity of the procedure. We run a treatment that show participants their own inference answers when they solve the corresponding forecast-revision problems, effectively reducing the two-step infer-then-LIE procedure to a one-step procedure of simply applying the LIE. The treatment, however, has little impact on the gap. Moreover, we confirm that participants are largely capable of applying the LIE correctly when solving a standalone expectation-formation problem. Taken together, these results suggest that participants do not appear to be following the infer-then-LIE 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 25%; meanwhile, two other behaviors that rarely appear in *Inference* become modal. Under the first mode, which represents 20% 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 (the state more consistent with the signal)—a simplifying heuristic that we term "exact representativeness."[2] The second mode, constituting 10% of the answers, is to report a forecast that equals the signal. That is, participants directly use the past realization as their expectation of the next outcome—a simplifying heuristic we term "naive extrapolation." Each of the three modal behaviors corresponds to participants

[2]This behavior also reflects 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).

using a different salient cue provided by the information environment—prior, outcome expectations (conditional on states), and 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 informational cues that appear similar to the variable elicited by that question. This explains the emergence of the two simplifying heuristics in *Forecast Revision*. Because the two conditional expectations of price growth appear more similar to expected price growth (elicited by forecast questions) than to firm quality (elicited by inference questions), they are more likely to be used as an anchor when participants make forecasts. Similarly, because realized price growth appears similar to expected price growth, it serves as a salient anchor when participants make forecasts, resulting in naive extrapolation.

To further test this similarity-based hypothesis, we run two additional treatments. In the first treatment, we rephrase *Forecast Revision* to *decrease* the similarity between the elicited variable and informational cues while keeping *Inference* relatively intact. Consistent with our hypothesis, exact representativeness and naive extrapolation become less prevalent in *Forecast Revision*, and the inference-forecast gap disappears. In the second treatment, we rephrase *Inference* to *increase* the similarity between elicited variable and information cues while keeping *Forecast Revision* relatively intact. This change makes exact representativeness and naive extrapolation more prevalent in *Inference* and reduces the inference-forecast gap. Overall, these treatments support the view that the similarity between informational cues in the environment and the belief-updating question can ultimately determine whether one underreacts or overreacts.

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).[3] 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.[4] 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.[5]

The finding of overreaction in forecast revisions provides experimental support for overreaction in survey expectations.[6] 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;

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

[4]A few belief-updating experiments using the bookbag-and-poker-chip design elicit beliefs of future draws conditional on the current draw. Moreno and Rosokha (2016), 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.

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

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

He and Kucinskas, 2020).[7] 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.[8] First, the explicit separation between states and outcomes makes it possible to design different questions targeting inference and forecast revision, respectively, thereby allowing us to pin down where a specific updating bias arises. Second, such a design allows us to separately identify different forms of overreaction, such as representativeness-based overreaction (Kahneman and Tversky, 1972; Bordalo et al., 2018) and mechanical extrapolation (Barberis et al., 2015, 2018). Indeed, both forms are prevalent in the data and contribute to the inference-forecast gap. Third, having a fully-specified DGP allows us to attribute biases in posterior beliefs to incorrect statistical reasoning rather than to misperceived DGPs. We also apply this design to show that overreaction in forecast revision generalizes to a setting in which signals and outcomes are of two different variables.

Overreaction in *Forecast Revision* is reminiscent of the hot-hand bias (Gilovich et al., 1985; Tversky and Gilovich, 1989; Suetens et al., 2016), which refers to the exaggeration of belief in an outcome after observing a long streak of the same outcomes. 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, 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). On the design level, we use explicit instructions and comprehension checks to make sure participants do not commit the hot-hand fallacy. Overall, it is unlikely that our results are driven by or a manifestation of the hot-hand bias.

[7] He and Kucinskas (2020) also finds that forecasts underreact to past observations of a different variable.

[8] 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).

The rest of the paper proceeds as follows. Section 2 outlines our experimental design. Section 3 shows 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 that shares the same information environment with an identical DGP and realized signal.

The *Baseline* treatment has five parts which are 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, and participants are provided with the conditional distributions of signals: signals of a good firm follow an i.i.d. normal distribution of $N(100, \sigma^2)$ and signals of a bad firm follow i.i.d. $N(0, \sigma^2)$.[9] Because good firms are more likely to have higher stock price growth than bad firms, a signal of high stock price growth (higher than 50) is diagnostic of the firm being good.

To sum up, in each round, the DGP is fully specified by two pieces of information: the prior distribution of states and the conditional distribution of signals. Both are presented to participants using figures and texts in a one-page display (see Figure 2 for an example), and we explain this interface with detailed instructions.[10] Table 2 summarizes the parameter values for the eight DGPs.

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

[10] Screenshots of the experimental interface can be found in the online appendix.

Table 1: Summary of variables elicited in each part of the experiment

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

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, each problem in any given part has a corresponding problem in each of the other four parts, which ensures that answers across parts are directly comparable. Unless mentioned otherwise, an observation is defined as a participant's answers to the five corresponding questions in the five parts.

The two main parts 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 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.[11] In *Forecast Revision*, participants instead report their updated expectations about the firm's stock price growth next month $\mathbb{E}(s_1|s_0)$. To ensure an apples-to-

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

There is a new pool of 20 firms.
The figure below describes the **stock price growth** of good firms and bad firms in any given month:
The green bar on top of each number is the chance (%) that a good firm's stock price grows by that number (in ¢) in any given month.
The orange bar on top of each number is the chance (%) that a bad firm's stock price grows by that number (in ¢) in any given month.

The pool of firms has the following composition.

Figure 2: An example of the interface for the DGP

apples 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*, they directly report their prior beliefs about the states $Pr(\theta)$ based on their knowledge about the DGP. Similarly, in *Forecast Prior*, they directly report their 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 $40\% \times 20 = 8$ good firms and 12 bad ones.[12] *Expectation Formation* is designed to test whether participants can correctly form expectations about the next signal when the states are distributed according to their own inference posteriors.

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 posteriors and (b) the *Expectation Formation* part comes after *Inference*. Hence, we are left with three orders of parts: 12345, 12534, and 34125.

After the five parts, the experiment ends with an unincentivized exit survey. At the end of the experiment, participants may receive a \$5 bonus payment, the chance of which depends on their answer in one randomly selected round through a quadratic rule.[13]

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

[13]If their answer in that round equals the rational benchmark according to standard probability theory, then they receive the bonus with certainty; otherwise, their chance of getting the bonus decreases quadratically in the difference between their answer and the rational benchmark (see (Hartzmark et al., 2021) for a similar incentive structure). If the

Building on the *Baseline* treatment, 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/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 the *Baseline* treatment 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 (LIE) 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 expectation of the forecast outcome s_1. If an *Inference* answer and its corresponding *Forecast Revision* answer satisfy this condition, then there is no discrepancy between these two types of belief-updating problems: Bayesian inference would then 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. Un-

answer is p and the rational benchmark is q (in % for the two *Inference* parts), then the chance of receiving the bonus is $\max\{0, (100 - (p - q)^2)\%\}$.

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

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-LIE* 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 LIE using the posteriors from the first step to obtain their expectations about the forecast outcome.

Since the correct implementation of the infer-then-LIE procedure satisfies the no inference-forecast gap condition, a gap can arise for two broad reasons. First, participants may consciously follow the infer-then-LIE procedure, but in doing so make errors or take shortcuts that bias their expectations, resulting in a gap. Second, it may be that participants do not use the infer-then-LIE procedure, but use alternative procedures in their forecast revisions.

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

[14]Moreover, because beliefs are equally incentivized across the two types of problems, rational tradeoff 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.[15]

2.4 Procedural details

We programmed our experiment using oTree (Chen et al., 2016). For *Baseline*, we recruited 202 participants through Prolific, an online platform designed for social science research.[16] Signals were framed as monthly revenue growth for 120 participants and as stock price growth for 82 participants. For 40 participants, questions in the forecast parts—namely Parts 3, 4, and 5 in Table 1—asked about expectations of the *last* signal ("stock price or revenue growth in the previous month") instead of the *next* signal. There was also some variation across participants in the order of parts: 72 participants went through the experiment in the order of 12345, 73 in the order of 12534, and 57 in the order of 34125. The participants, on average, spent about 30 minutes on the experiment and earned a payment of $7.15, $5 of which was the base payment.

2.5 Other treatments

In addition to *Baseline*, we also implemented several other treatments that 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.

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

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

Table 3: Overview of additional treatments

Treatment	Section	Difference from *Baseline*
Cross-variable Forecast	3.2	Forecast outcome is a different variable: $= 100$ if $\theta = G$, $= 0$ 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
Similarity I	5.1	Forecast outcome is a different variable: =Up if $\theta = G$, =Down if $\theta = B$; forecast questions ask about full distributions
Similarity II	5.2	State variable (profitability) = mean of signal / forecast outcome (profits); inference questions ask about the expectation of the state

3 Evidence for the Inference-Forecast Gap

In this section, we present results from our experiment to compare belief-updating in inference and forecast-revision problems. This comparison is carried out using three methods of analysis. First, we classify answers into categories of *Near-rational*, *Overreact*, and *Underreact*, and compare the distributions of these three categories amongst inference and forecast-revision problems. Second, we compare the average belief movements from the priors to the posteriors in these two types of updating tasks. Third, we compare the distributions of individual answers and identify differences in modal behaviors. If the no inference-forecast gap condition in Equation (2) is met, then results from the two types of updating problems should exhibit identical patterns in all three kinds of analysis. Any systematic differences would imply the existence of an inference-forecast gap.

3.1 Aggregate patterns

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

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

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

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

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

We classify answers in *Inference* and *Forecast Revision* by how they compare to the rational benchmarks. An answer is classified as *Near-rational* if its difference from the rational benchmark is no more than 2.5.[17] To introduce the categories of *Underreact* and *Overreact*, 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:

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

It is straightforward from equations (3) and (4) that rational updates in any two corresponding inference and forecast-revision problems are identical. We classify an answer as *Overreact* if its update is larger than the rational update by more than 2.5 and as *Underreact* if its update is smaller than the rational update by more than 2.5. We do not classify answers when $s_0 = 50$; that is, when the signal is uninformative.

Table 4 shows the aggregate patterns in the *Baseline* treatment (excluding observations with a realized signal of 50). 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, 59.8% imply underreaction, 25.2% imply overreaction, and 15% are considered *Near-rational*. These patterns, however, flip in *Forecast Revision*: 49.5% of the answers indicate overreaction to new information—higher than

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

Table 4: Aggregate patterns in *Baseline*

N=202, Obs=1480	Classification			Update
	Underreact	*Near-rational*	*Overreact*	Mean (s.e.)
Inference	59.8%	15.0%	25.2%	15.1 (0.8)
Forecast Revision	43.1%	7.4%	49.5%	29.9 (2.3)
Rational				23.4 (0.3)

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

the 43.1% classified as underreaction.

The last column of Table 4 demonstrates that the inference-forecast gap also shows up in average updates. In *Inference*, the average update across all answers is 15.1, significantly smaller than the average rational update of 23.4. By contrast, in *Forecast Revision*, participants update too much, leading to an average update of 29.9. Column (1) of Table A6 confirms the inference-forecast gap in updates in a regression.

The inference-forecast gap is highly robust in various cuts of the data (see Appendix A for detailed results). In a more "reasonable" subsample which only includes observations with a forecast-revision answer within $[0, 100]$ and updates that are nonnegative, forecast-revision answers no longer exhibit overreaction on average, but the inference-forecast gap remains highly significant. Moreover, the gap is present under all eight DGPs, which entail different priors and signal distributions. The gap increases in signal strength but exists even for the weakest signals. Our result also persists in a subsample that excludes observations with incorrect reported prior beliefs. In addition, the order of experimental parts, whether we frame the signals and outcomes as stock price growth or revenue growth, and participant characteristics have no qualitative impacts on the inference-forecast gap. One framing variation that has a significant impact on the magnitude of the gap is the variation in the timing of outcome realization. The gap is half as large when the outcome is framed as "stock price / revenue growth of the last month" as when it is framed as the

outcome of the next month (see Table A9). This result suggests that beliefs about unrealized events are more responsive to signals than beliefs about realized events, which could partially account for the inference-forecast gap.[18]

3.2 *Cross-variable Forecast* 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 the *Baseline* treatment. These results demonstrate the prevalence of the inference-forecast gap in various environments and help rule out several potential explanations for its emergence.

The forecast outcome in *Baseline* is the next signal, which is identical to the realized signal both in name and in distribution. We change this feature in an additional treatment (N=100) called *Cross-variable Forecast* in which the forecast outcome is a variable different from the signal. Specifically, the new outcome variable is framed as revenue growth when the signal is stock price growth, and vice versa. We also design the outcome variable to have a degenerate distribution conditional on the state: It is 100 for sure in the Good state and 0 for sure in the Bad state. Thus, the outcome is different from the signal in both name and distribution and is similar to the state in distribution. Nevertheless, under this alternative DGP, the no inference-forecast gap condition remains the same as before: the forecast-revision answer equals the corresponding inference answer (minus the % sign).

Table 5 shows the results from *Cross-variable Forecast*. The inference-forecast gap, compared to that in *Baseline*, becomes even greater in magnitude. For example, only 21.1% of the inference posteriors are classified as *Overreact*, while 50.7% of the forecast-revision answers are. Table A10 further shows, in a regression analysis, that statistically the gap is highly significant. In Table B4, we show that the distribution of behavioral modes in *Cross-variable Forecast* is also similar to that

[18]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) find that people prefer earlier resolution of uncertainty for realized events than for unrealized ones.

Table 5: Aggregate patterns in *Cross-variable Forecast*

N=100, Obs=748	Classification			Update
	Underreact	*Near-rational*	*Overreact*	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 *Underreact*, *Near-rational*, and *Overreact*. The last column shows average belief movements in the signal direction from the (objective) priors and their rational benchmark. Observations with signal equal 50 are excluded. Standard errors are clustered by subject.

in *Baseline*.

The results from *Cross-variable Forecast* help address four issues. First, they rule out the possibility that the inference-forecast gap is driven by the common name and distribution shared by the signal and the outcome. According to this explanation, these common features of the signal and the outcome lead participants to perceive the signal as more informative and therefore to overreact to it in *Forecast Revision*. The results in *Cross-variable Forecast* clearly demonstrate that participants still overreact to signals when they are asked to make predictions about a different variable.

Second, the fact that the state fully determines the outcome indicates that the inference-forecast gap is not due to the difference in distribution between the state and the outcome. Under this design, signals are equally diagnostic about the state and the outcome, thereby ruling out any explanations based on differential diagnosticity between inference and forecast-revision problems.

Third, the presence of overreaction in *Cross-variable Forecast* suggests that the overreaction in forecast-revision problems is not driven by misperceived signal autocorrelation. Because the forecast outcome is different from the signal and fully determined by the state, perception of signal autocorrelation is irrelevant to the expectation formation of the future outcome. This further differentiates our results from the hot-hand bias (Gilovich et al., 1985; Tversky and Gilovich, 1989; Suetens et al., 2016) and from overreaction in univariate forecasts (Hey, 1994; Frydman and Nave,

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%

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

2017; Afrouzi et al., 2020) in which exaggerated autocorrelation is a key driving force.

Fourth, *Cross-variable Forecast* broadens the external relevance of the inference-forecast gap. In many empirical settings, the forecaster's information set is not limited to past observations of the variables to be forecasted but also includes past observations of other relevant variables. The results in *Cross-variable Forecast* suggest that the inference-forecast gap can be an explanation for overreactions in these settings as well (e.g., Bordalo et al., 2020; Roth and Wohlfart, 2020).

3.3 *Binary Signal* treatment

We implement a treatment (N=140) in which the signal s_t follows a binary distribution. The signal, framed as the direction of the firm's stock price movement, is either up or down, and the probability of an upward movement is higher if the firm's state is Good. The parameters for the DGPs are listed in Table 6. In the forecast-revision part of the treatment, the problem asks about the probability distribution $Pr(s_1)$ (instead of the outcome expectation $\mathbb{E}(s_1)$).

As in the *Baseline* treatment, the no inference-forecast gap condition for this treatment 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

Table 7: Aggregate patterns in *Binary Signal*

N=140, Obs=1120	Classification			Update
	Underreact	*Near-rational*	*Overreact*	Mean (s.e.)
Inference	61%	20.1%	18.9%	11 (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 *Underreact*, *Near-rational*, and *Overreact*. 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 signal equal 50 are excluded. Standard errors are clustered by subject.

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. However, one advantage of the *Binary Signal* treatment is that it is closer to the common design in the bookbag-and-poker-chip paradigm.

In *Binary Signal*, the three categories—*Near-rational*, *Underreact*, and *Overreact*—are defined in the same way as in the *Baseline* treatment, except that the categories for forecast-revision answers are defined based on their *normalized* updates. Table 7 reports the results from the *Binary Signal* treatment. As in *Baseline*, more answers are classified as *Overreact* in *Forecast Revision* than in *Inference*, and the average update in the former part is also larger. (Table A11 shows in a regression that the gap in updates is significant at the 10% level.) However, answers in *Forecast*

Revision do not exhibit overreaction on average. The modal behaviors are also similar to those in the *Baseline* treatment (see Table B5). Non-updates are prevalent in both *Inference* and *Forecast Revision*, making up 27.1% and 19.8% of answers in those two parts, respectively. In *Forecast Revision*, 17.4% of the answers equal the outcome probability of the representative state, which constitutes the behavioral mode of exact representativeness.

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 distribution of the outcome instead of its expected value.

4 Decision Procedures

4.1 Implementation errors or nonstandard procedures?

After documenting the existence of the inference-forecast gap, 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, in answering a forecast-revision question, correctly implement the infer-then-LIE procedure by: (a) first updating their beliefs about the states as in the corresponding inference problem and (b) then applying the LIE to form expectations about the forecast outcome. The evidence we have documented so far on the inference-forecast gap clearly rejects the correct implementation of this procedure, prompting us to look for alternative explanations.

One possible explanation for the inference-forecast gap is that participants intend to follow the infer-then-LIE procedure when revising forecasts, but make errors or take shortcuts because the procedure is complex. For instance, a decision-maker may be capable of forming probabilistic beliefs about the states when making inference is the only task. But when implementing the two-step infer-then-LIE procedure for the forecast-revision problem, she may have only enough cognitive bandwidth to form a binary belief ("the firm is good" or "the firm is bad") in the first step. This

Table 8: Aggregate patterns in *Nudge*

N=99, Obs=715	Classification			Update
	Underreact	*Near-rational*	*Overreact*	Mean (s.e.)
Inference	70.6%	10.2%	19.2%	10.3 (1.3)
Forecast Revision	42.2%	6.7%	51%	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 *Underreact*, *Near-rational*, and *Overreact*. 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 signal equal to 50 are excluded. Standard errors are clustered by subject.

error can lead to overreacting behaviors that look like exact representativeness.

We run an additional treatment, *Nudge*, with 99 participants to test the above hypothesis. In parts that provide signals, after observing the realized signals, participants are first asked to report their beliefs about the state and then, while their answers are still displayed on the screen, they are asked to report their expectations about the next signal.[19] With this design, from the point of view of a participant intending to follow the infer-then-LIE procedure, a forecast-revision problem is made no more complex than a standalone expectation-formation problem. Indeed, one only need to multiply the inference posterior by 100 to complete the infer-then-LIE procedure.[20] According to the hypothesis above, this reduction in complexity should mitigate any implementation errors in the procedure and reduce the inference-forecast gap.

Table 8 shows the aggregate patterns in *Nudge*. In this treatment, participants overwhelmingly

[19]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.

[20]In fact, because answers to the *Inference* problems are given in the unit of percentage, the infer-then-LIE procedure implies that participants should type in exactly the same number in the corresponding *Forecast Revision* problems.

underreact in *Inference* and on average overreact in *Forecast Revision*. In fact, the inference-forecast gap in *Nudge* is even larger than in *Baseline*, according to the regression analysis in Table A10. Table B6 further examines the modal behaviors in *Nudge*. The fraction of non-updates in *Inference* is 53.4%, a notable increase from the 29.9% 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 11.3% from the 5.3% in the *Baseline* treatment, suggesting that the *Nudge* treatment induces a greater tendency to give internally consistent answers to the two types of updating questions. However, this small increase does not have material impact in the aggregate. Taken together, displaying the inference answer when participants revise their forecasts does not change the overall pattern of the inference-forecast gap.

How can one explain the ineffectiveness of the *Nudge* treatment? One possibility is that while it indeed makes the infer-then-LIE procedure no more complex than solving a standalone expectation-formation problem, even the latter is too complex for participants and the resulting errors lead to overreaction. To test this possibility, in another part of the *Nudge* treatment called *Expectation Formation*, we ask participants to report their beliefs about the state and then their expectations of the next signal *without* showing them any signal realization. In addition, for each participant, we set the distribution over states in an expectation-formation problem to match the posterior belief the participant 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 $40\% \times 20 = 8$ good firms and 12 bad ones. This design enables us to directly quantify how much of the inference-forecast gap in *Nudge* can be attributed to mistakes in expectation formation.

Figure C2 in Appendix C shows the average deviations from LIE in the expectation-formation problems by the prior probability of the Good state; the deviations are small across the board. Moreover, in the last row of Table 8, we classify expectation-formation answers and calculate

their updates by treating them in the same way as their corresponding forecast-revision answers. Specifically, 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. Comparing the average updates in the inference, forecast-revision, and expectation-formation problems, we find that mistakes in expectation formation can account for only 18% of the inference-forecast gap. These results indicate that mistakes in standalone expectation-formation problems do not explain the null effect of the *Nudge* treatment on the inference-forecast gap.

Taken together, results from the *Nudge* treatment reject the hypothesis that the inference-forecast gap stems from complexity-induced errors or shortcuts when participants try to implement the infer-then-LIE procedure in forecast-revision problems. Rather, the gap is likely a result of the use of other procedures altogether.

4.2 Alternative decision procedures

What alternative decision procedures do participants use? To answer this question, we examine the distributions of posterior beliefs and look for modal behaviors that could shed light on the underlying decision procedures. To illustrate, Figure 3 plots the answers against the realized signals for problems with symmetric objective priors in *Inference* and *Forecast Revision*.[21] Several behavioral modes appear salient in the plots. In *Inference*, a large fraction of answers equals the 50-50 prior. 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. However, two other modes that are exclusive to forecast-revision problems emerge. First, a large number of forecast-revision answers cluster at 100 when $s_0 > 50$ and 0 when $s_0 < 50$. Participants who give these answers behave as

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

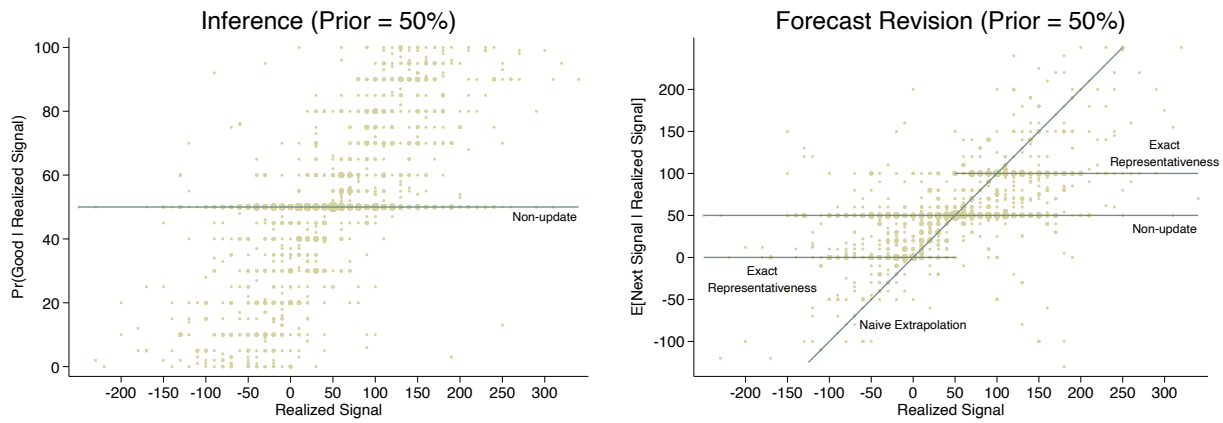

Figure 3: Scatterplots of answers against realized signals: subsample with symmetric priors

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.

if they were certain about being in the representative state (the state consistent with the direction of the signal's 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).[22]

Second, a smaller yet still significant fraction of forecast-revision answers are anchored at the face value of the realized signal.[23] We term this behavior "naive extrapolation" because it suggests a particular form of extrapolative expectation formation (Barberis et al., 2015, 2018; Liao et al., 2021).[24] The face value of the realized signal is among the top three common answers for 19 out of 53 values of the realized signal. 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 all inference and

[22] An alternative interpretation of this modal behavior is that participants base their expectations solely on the ex-post more likely state (Mullainathan, 2002), which can differ from the representative state when the prior is asymmetric. We can differentiate the two interpretations by examining problems with an asymmetric prior. In Appendix B, we study the distributions of forecast-revision answers under asymmetric priors and find evidence supporting the representativeness interpretation. However, this result should be interpreted as only suggestive due to a small sample size.

[23] 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.

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

Table 9: Modes of behavior in *Baseline*

Mode	Criterion for answer	*Inference*	*Forecast Revision*
Non-update	= prior	29.9%	25.1%
Exact Representativeness	= 100 if $s_0 > 50$, = 0 if $s_0 < 50$	3.9%	20.1%
Naive Extrapolation	$= s_0$	3.3%	10.3%
No Inference-Forecast Gap (excluding the other modes)	inference = forecast revision	3.6%	
Unclassified		59.8%	43.9%
Observations		1480	1480

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 signal equal to 50 are excluded.

forecast-revision problems. Confirming the patterns in the scatterplots, non-updates are widespread in both types of problems, making up 29.9% and 25.1% of the answers in *Inference* and *Forecast Revision*, respectively. The other two behavioral modes, exact representativeness and naive extrapolation, appear almost exclusively in *Forecast Revision*, making up 20.1% and 10.3% of the answers, respectively. In comparison, observations that meet the no inference-forecast gap condition and are not in any of the three behavioral modes constitute only 3.6% of the answers. We conduct further analysis in Appendix B. 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. Based on this participant-part–level classification, we also find 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.

In Table A7, we show that the difference in modal behaviors is an important driver of the inference-forecast gap in the aggregate. The inference-forecast gap shrinks by 36% when we

exclude from the full sample observations with at least one answer in one of the two distinctive behavioral modes in *Forecast Revision*—exact representativeness and naive extrapolation. In the 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.

All the behavioral modes, albeit capturing different answers, share one common feature—each of them solely relies on one salient cue in the information environment. Specifically, non-updates, exact representativeness, and naive extrapolation base the answer entirely on the prior, the outcome expectations conditional on the states, and the realized signal, respectively. This feature of the behavioral modes reflects the use of simplifying heuristics because focusing on a few cues instead of properly aggregating all relevant information is 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 inference and forecast-revision tasks, even when the information environments are identical. 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 cues in the information environment and the belief-updating question. Specifically, when answering a belief-updating question, people are more likely to rely on salient cues that appear similar to what the question asks for. For example, in our *Baseline* treatment, the forecast-revision question asks for the expected stock price growth conditional on the realized signal. The subject matter, expected price growth conditional on the 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 induces participants to rely on ex-

pected price growth conditional on the representative state as a cue in making forecasts, resulting in exact representativeness. In contrast, the inference question asks about the conditional probability distribution over the states, which is less similar to expected price growth conditional on the representative state. As a result, exact representativeness is rarely observed in inference tasks.

By the same logic, the realized signal, which is the cue used in naive extrapolation, is a monetary measure of the firm's performance, just like the subject matter of the forecast-revision question. Therefore, participants employ the heuristic of naive extrapolation in forecast-revision tasks more than in inference tasks, where the cue appears less similar to the question. The prior belief over the states and the prior outcome expectation are equally similar to their posterior counterparts. Hence, non-updates, the heuristic that focuses on the prior, are prevalent in inference and forecast-revision tasks to roughly the same extent. These arguments are summarized in Table 10.

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

Cue	Inference Pr(state | realized stock price)	Forecast Revision E(stock price | realized stock price)	Behavior
E(stock price | representative state)	Not similar	Similar	Exact Representativeness
Realized stock price	Not similar	Similar	Naive Extrapolation
E(stock price)		Similar	Non-update
Pr(state)	Similar		

To test this hypothesis, we design two treatments which manipulate the similarity between cues and belief-updating questions by varying the framing of variables and questions. If our hypothesis is correct, then the changes in cue-question similarity will affect the heuristics participants employ, which in turn will affect the modal behaviors and the aggregate inference-forecast gap.

5.1 *Similarity* treatment I

In the first treatment that tests our hypothesis, we change the forecast-revision question to make it *less* similar to two salient cues. Specifically, the forecast-revision question in this treatment asks

about the probability that the firm's revenue will go up next month conditional on the realized stock price growth of the current month. Moreover, participants are informed that a firm's revenue goes up *if and only if* the state is Good. The inference question remains intact.

Table 11 summarizes how this treatment affects the cue-question similarity. Note first that the expected stock price conditional on the states are less similar to the forecast-revision question in this treatment, which asks how likely the revenue will go up. Second, the probabilities that the revenue goes up conditional on the states (100 or 0) are arguably not as salient as the other cues in the information environment. As a result, fewer forecast-revision answers should follow the heuristic of exact representativeness. Moreover, the forecast question also becomes less similar to the realized signal (how much the stock price grew in the current month) compared to the *Baseline* treatment. This should make naive extrapolation less prevalent.

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

Cue	Inference Pr(state | realized stock price)	Forecast Revision Pr(revenue goes up | realized stock price)	Behavior
Pr(revenue goes up | representative state)	Not salient		Exact Representativeness
E(stock price | representative state)	Not similar	Not similar	
Realized stock price	Not similar	Not similar	Naive Extrapolation
E(stock price)		Similar	Non-update
Pr(state)	Similar		

The first two columns of Table 13 show the modal answers in the first *Similarity* treatment. Exact representativeness and naive extrapolation are much less prevalent in *Forecast Revision* of this treatment compared to the *Baseline*. This change in modal behaviors supports our hypothesis that when a cue becomes less salient or less similar to the question being asked, the heuristics that people use are less likely to rely on this cue. Another pattern is that the fraction of answers that satisfy the no inference-forecast gap condition and are not in the three behavioral modes increases from 3.6% in *Baseline* to 11.8% in the first *Similarity* treatment. One possible explanation for this is that the design makes it easier for some participants to recognize the tight conceptual connection between inference questions and forecast-revision questions.

Table 12: Similarity between belief-updating questions and salient cues in *Similarity II*

Cue	Inference E(profitability | realized profit)	Forecast Revision E(profit | realized profit)	Behavior
E(profit | representative state)	Similar	Similar	Exact Representativeness
Realized profit	Similar	Similar	Naive Extrapolation
E(profit)		Similar	Non-update
E(profitability)	Similar		

The change in modal behavior also alters the aggregate pattern of the inference-forecast gap. Table 14 shows that the inference-forecast gap almost completely vanishes in the first *Similarity* treatment, and we obtain the familiar underreaction pattern in the forecast-revision problems.

5.2 *Similarity* treatment II

In the second treatment that tests the similarity hypothesis, we reframe the variables and the inference question to *increase* the similarity between the inference question and two salient cues. Specifically, the signal and outcome variables are framed as the firm's profits of the current month and of the next month. The state variable is framed as a firm's profitability, which is 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 in *Baseline*. The inference question asks about the expected profitability of a firm after the realization of the current month's profit. The forecast-revision question asks about the expected profit of the next month conditional on the same event.

Table 12 summarizes how this treatment affects the cue-question similarity. Note that the inference question asks about expected profitability. This subject matter is a monetary measure of the firm's performance, similar to expected profits conditional on the states and the realized profits, two salient cues in the information environment. According to our hypothesis, we expect that the two heuristics based on these two cues, exact representativeness and naive extrapolation, will be more prevalent in inference problems of this treatment than in the *Baseline*.

Table 13: Modes of behavior in the two *Similarity* treatments

	Similarity I		Similarity II	
Mode	*Inference*	*Forecast Revision*	*Inference*	*Forecast Revision*
Non-update	31.7%	30.8%	33.3%	31.2%
Exact Representativeness	9.0%	13.9%	17.9%	19.9%
Naive Extrapolation	3.9%	3.6%	22.6%	31.4%
No Inference-Forecast Gap (excluding the other modes)	11.8%		3.8%	
Unclassified	45.2%	41.5%	25.8%	17.4%
Observations	467	467	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 signal equal to 50 are excluded.

Table 14: Aggregate patterns in the *Similarity* treatments

Treatment		Classification			Update
		Underreact	*Near-rational*	*Overreact*	Mean (s.e.)
Similarity I	*Inference*	64.7%	12.2%	23.1%	14.3 (1.6)
(N = 60)	*Forecast Revision*	62.1%	12.8%	25.1%	13.6 (1.8)
(Obs = 467)	Rational				23.1 (.6)
Similarity II	*Inference*	48.6%	4.8%	46.6%	30.1 (4.4)
(N = 60)	*Forecast Revision*	38.0%	2.9%	59.0%	41.6 (5)
(Obs = 442)	Rational				24.4 (.6)

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

The results of this treatment support the similarity hypothesis. Table 13 shows that exact representativeness and naive extrapolation become modal behaviors in the inference problems of the second *Similarity* treatment. This is in stark contrast with the *Baseline* treatment where these two behaviors are almost non-existent in *Inference*. This treatment also generates a different aggregate pattern from *Baseline*: the numbers of underreacting and overreacting answers are approximately the same in *Inference*, and the average inference update is even on the overreaction side (see Table 14). Table A10 shows in a regression that the inference-forecast gap is smaller compared to *Baseline* but 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 entirety of the inference-forecast gap.

6 Concluding Remarks

In this paper, we present new experimental evidence to show that people overreact more to new information when they revise forecasts about future outcomes than when they make inferences about underlying states. 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. We further examine the underlying mechanism and show that the gap does not stem from errors participants make when they try to implement a standard decision procedure for forecast revision. Instead, the discrepancy is largely driven by the use of different simplifying heuristics in the two types of updating problems. To explain these different heuristics, we build our hypothesis on the literature on salience and memory retrieval. Studies in this literature have shown that people's answers to a question are more likely to rely on cues that appear similar to what the question asks for. In two treatments, we vary the similarity between updating questions and the salient cues in the information environment. Consistent with our hypothesis, these treatment variations affect the heuristics being used, which in turn affects the aggregate inference-forecast gap. Overall, our results open up a new direction for the study of belief updating—biases in posterior

beliefs can depend on the type of updating question even when the information environment is fixed. This is because different updating questions trigger different decision procedures.

Recent field studies find widespread overreaction in survey forecasts of variables such as stock returns, macroeconomic indicators, firm performances, housing prices, and job offers. This is in stark contrast with the prevalent underreaction documented in laboratory experiments on inference. While our experimental evidence is consistent with both sets of facts, we do not claim that the inference-forecast gap explains the entire discrepancy between these two literatures. After all, field settings are different from the laboratory in many other aspects that could be driving overreaction in survey forecasts. First, the DGP can be much more complex in reality than in simple experimental settings. The underlying state may be time-variant and the forecast outcome may be correlated with past signals even conditional on the state. The DGP itself may even be unknown. Second, survey takers may come from a different pool than experimental participants. For instance, many financial survey participants are professional forecasters or investors who are more likely to possess a good understanding of the financial market.

Despite the caveats, we believe our evidence can still speak—at least partially—to what is going on in the field for the following reasons. First, with more complex DGPs in reality, it could be all the more likely that people revise their forecasts using heuristics that are detached from their beliefs about fundamentals. Second, while survey takers may be more sophisticated, many market participants are households and closer to the participants we study. It is also worth noting that even professionals' forecasts have subjective inputs (Stark, 2013) and are highly correlated with household expectations (Greenwood and Shleifer, 2014).[25]

References

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

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

P. Andre, C. Pizzinelli, C. Roth, and J. Wohlfart. Subjective models of the macroeconomy: Evidence from experts and representative samples. *Available at SSRN 3355356*, 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*, 2021.

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.

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. Technical report, 2021a.

P. Bordalo, N. Gennaioli, A. Shleifer, and S. J. Terry. Real credit cycles. Technical report, 2021b.

K. Burdett and T. Vishwanath. Declining reservation wages and learning. *The 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, July 2015. doi: 10.1257/aer.20140117. URL https://www.aeaweb.org/articles?id=10.1257/aer.20140117.

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. Technical report, National Bureau of Economic Research, 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. *The Quarterly Journal of Economics*, 135(3):1363–1398, 2020.

B. Enke and T. Graeber. Cognitive uncertainty. 2020.

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

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

L. G. Epstein, Y. Halevy, et al. Hard-to-interpret signals. 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. 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. *The Quarterly Journal of Economics*, 129(4):1661–1710, 2014.

N. Gennaioli and A. Shleifer. What comes to mind? *The 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. *Available at SSRN 3658112*, 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. 2021.

S. He and S. Kucinskas. Expectation formation with correlated variables. *Available at SSRN 3450207*, 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. 2021.

Y. Liang. Learning from unknown information sources. *Available at SSRN 3314789*, 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. *The 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. Technical report, Working Paper, Harvard University, 2002.

S. Mullainathan, J. Schwartzstein, and A. Shleifer. Coarse thinking and persuasion. *The 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.

C. Roth and J. Wohlfart. How do expectations about the macroeconomy affect personal expectations and behavior? *Review of Economics and Statistics*, 102(4):731–748, 2020.

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 marked 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 observations indicate that the participants' reactions to signals are in the wrong direction.

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

N=202, Obs=978	Classification			Update
	Underreact	*Near-rational*	*Overreact*	Mean (s.e.)
Inference	55.8%	17.3%	26.9%	17.7 (1)
Forecast Revision	45.6%	10.1%	44.3%	23.1 (1.4)
Rational				23.5 (0.4)

Notes: The first three columns present the percentages of answers classified as *Underreact*, *Near-rational*, and *Overreact*. The last column shows average belief movements in the signal direction from the (objective) priors and their rational benchmark. Observations with 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 subject.

Table A1 shows the results of 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 eight problems (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 decision-maker 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 exists 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 a 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 parts. Table A5 compares the gap across different orders and shows that there is a large and statistically significant gap for all three. 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
		Underreact	*Near-rational*	*Overreact*	Mean (s.e.)
$Pr(G)$ = 50%	*Inference*	69.3%	18.8%	12%	20 (1.5)
σ = 50	*Forecast Revision*	45.8%	13.5%	40.6%	30.7 (2.9)
(Obs = 192)	Rational				36.8 (0.9)
$Pr(G)$ = 50%	*Inference*	64.5%	18.8%	16.7%	17.9 (1.5)
σ = 60	*Forecast Revision*	48.4%	7%	44.6%	28.9 (3.4)
(Obs = 186)	Rational				32.4 (1)
$Pr(G)$ = 50%	*Inference*	64.7%	12.6%	22.6%	15.6 (1.4)
σ = 70	*Forecast Revision*	43.2%	7.9%	48.9%	28.1 (3.2)
(Obs = 190)	Rational				26.7 (0.9)
$Pr(G)$ = 50%	*Inference*	65.1%	11.6%	23.3%	12.5 (1.4)
σ = 80	*Forecast Revision*	45%	5.3%	49.7%	29.5 (3.6)
(Obs = 189)	Rational				22.8 (0.9)
$Pr(G)$ = 50%	*Inference*	50.5%	17.9%	31.6%	17 (1.3)
σ = 90	*Forecast Revision*	40.5%	5.8%	53.7%	33.9 (3.8)
(Obs = 190)	Rational				21.2 (0.9)
$Pr(G)$ = 50%	*Inference*	52.6%	15.1%	32.3%	13.7 (1.4)
σ = 100	*Forecast Revision*	36.5%	7.8%	55.7%	33.6 (3.6)
(Obs = 192)	Rational				19.7 (0.9)
$Pr(G)$ = 80%	*Inference*	55.7%	10.9%	33.3%	12 (1.8)
σ = 100	*Forecast Revision*	44.8%	3.4%	51.7%	27.2 (4.6)
(Obs = 174)	Rational				13 (0.8)
$Pr(G)$ = 20%	*Inference*	55.1%	13.8%	31.1%	11 (2)
σ = 100	*Forecast Revision*	40.7%	7.8%	51.5%	26.3 (4.2)
(Obs = 167)	Rational				12.7 (0.8)

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

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

Signal Strength		Classification			Update
		Underreact	*Near-rational*	*Overreact*	Mean (s.e.)
Weakest	*Inference*	50.3%	21.2%	28.6%	4.9 (1)
(Obs = 189)	*Forecast Revision*	49.7%	13.8%	36.5%	10.2 (1.8)
	Rational				6.3 (0.2)
Weak	*Inference*	64.7%	12.7%	22.6%	8.8 (1.1)
(Obs = 252)	*Forecast Revision*	42.1%	5.6%	52.4%	20.5 (2.3)
	Rational				16 (0.2)
Medium	*Inference*	60.9%	7.9%	31.2%	15.6 (1.4)
(Obs = 202)	*Forecast Revision*	41.1%	4%	55%	31.4 (3.3)
	Rational				25.1 (0.2)
Strong	*Inference*	64.2%	11.2%	24.7%	21.1 (1.6)
(Obs = 215)	*Forecast Revision*	36.7%	4.2%	59.1%	46.4 (4.9)
	Rational				34.3 (0.2)
Strongest	*Inference*	63%	24.2%	12.8%	26.8 (1.6)
(Obs = 281)	*Forecast Revision*	46.3%	11.7%	42%	41.5 (4.2)
	Rational				44.9 (0.2)

Notes: The first three columns present the percentages of answers classified as *Underreact*, *Near-rational*, and *Overreact*. The last column shows average belief movements in the signal direction from the (objective) priors and their rational benchmark. Observations with 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 subject.

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

N=202, Obs=1095	Classification			Update
	Underreact	*Near-rational*	*Overreact*	Mean (s.e.)
Inference	57.7%	17.2%	25.1%	15.6 (1)
Forecast Revision	45.8%	8.8%	45.5%	26.3 (2.5)
Rational				23.8 (.4)

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

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

		Classification			Update
		Underreact	*Near-rational*	*Overreact*	Mean (s.e.)
Order: 12345	*Inference*	58.3%	16.1%	25.6%	14.9 (1.3)
(N = 72)	*Forecast Revision*	41%	8.7%	50.3%	30.1 (3.6)
(Obs = 527)	Rational				23.3 (0.5)
Order: 12534	*Inference*	58.6%	16.4%	25%	15.2 (1.3)
(N = 73)	*Forecast Revision*	42.9%	5.8%	51.2%	33 (3.8)
(Obs = 531)	Rational				23.5 (0.5)
Order: 34125	*Inference*	63.3%	11.8%	24.9%	15 (1.9)
(N = 57)	*Forecast Revision*	46%	7.6%	46.4%	25.5 (4.6)
(Obs = 422)	Rational				23.5 (0.6)

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

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 the same gap. 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 5% level. Table A9 show these results in regressions.

A.6 Regression analysis

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

	Update		
	Full sample	"Reasonable" updates	Correct priors
	(1)	(2)	(3)
Forecast Revision	14.801***	5.398***	10.642***
	(2.429)	(1.403)	(2.683)
Rational Update	1.012***	0.561***	0.923***
	(0.078)	(0.049)	(0.077)
Problem FE	Yes	Yes	Yes
Subject FE	Yes	Yes	Yes
Observations	2960	1956	2190
R^2	0.339	0.474	0.366

Notes: *, **, and *** indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively. Standard errors are clustered by subject. 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, 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 (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 *Inference Prior* or *Forecast Prior*.

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

	Update	
	Full sample & exclude two modes	"Reasonable" updates & exclude two modes
	(1)	(2)
Forecast Revision	9.537***	-2.980**
	(3.134)	(1.226)
Rational Update	0.927***	0.394***
	(0.103)	(0.061)
Problem FE	Yes	Yes
Subject FE	Yes	Yes
Observations	2006	1228
R^2	0.354	0.519

Notes: *, **, and *** indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively. Standard errors are clustered by subject. 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, 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	21.464***
	(4.658)
Male × Forecast Revision	-1.579
	(4.783)
College × Forecast Revision	2.835
	(4.690)
Investor × Forecast Revision	-1.445
	(5.123)
Familiar with Stats × Forecast Revision	-2.290
	(4.640)
Familiar with Econ × Forecast Revision	-9.176*
	(5.314)
High Comprehension × Forecast Revision	-9.705**
	(4.540)
Male	2.006
	(1.689)
College	-1.348
	(1.852)
Investor	3.548*
	(1.958)
Familiar with Stats	3.157
	(1.978)
Familiar with Econ	-3.139
	(2.326)
High Comprehension	5.006**
	(1.925)
Rational Update	0.987***
	(0.074)
Problem FE	Yes
Observations	2960
R^2	0.149

Notes: *, **, and *** indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively. Standard errors are clustered by subject. This table presents results for our baseline treatment. Each observation corresponds either to an inference answer or a forecast-0revision answer. We define the dependent variable, *Update*, as the answer minus the (objective) prior if the signal is greater than 50, 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. We define *Male* as 1 if the subject indicates their gender as Male; the base group is thus Female or Others. We define *College* as 1 if the subject has a bachelor's or postgraduate degree. We define *Investor* as 1 if the subject indicates that they have investments in stocks or mutual funds. We define *Familiar with Stats* as 1 if the subject indicates that they are familiar with probability theory and statistics. We define *Familiar with Econ* as 1 if the subject indicates that they are familiar with economics or finance. We define *High Comprehension* as 1 if the subject correctly answers all 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	14.858***	
	(3.465)	
Revenue × Forecast Revision	14.761***	
	(3.161)	
Revenue	4.316**	
	(1.726)	
Next × Forecast Revision		15.998***
		(2.651)
Last × Forecast Revision		9.779**
		(4.875)
Last		1.804
		(2.380)
Rational Update	0.991***	0.994***
	(0.075)	(0.074)
Problem FE	Yes	Yes
Observations	2960	2960
R^2	0.138	0.136

Notes: *, **, and *** indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively. Standard errors are clustered by subject. 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, 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 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	14.801***
	(2.341)
Cross-variable × Forecast Revision	19.198***
	(3.304)
Nudge × Forecast Revision	18.640***
	(2.962)
Similarity I × Forecast Revision	-0.665
	(1.642)
Similarity II × Forecast Revision	11.559***
	(3.526)
Cross-variable	-1.167
	(1.555)
Nudge	-4.218***
	(1.552)
Similarity I	-0.564
	(1.783)
Similarity II	14.143***
	(4.109)
Rational Update	0.942***
	(0.051)
Problem FE	Yes
Observations	7704
R^2	0.159

Notes: *, **, and *** indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively. Standard errors are clustered by subject. In this table, we pool the data from our *Baseline* treatment, *Cross-variable Forecast* treatment, *Nudge* treatment, and *Obvious Connection* treatment. Each observation corresponds either to an inference posterior or an extrapolation posterior. 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, 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.

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 subject. This table presents results for the *Binary Signal* treatment. Each observation corresponds either to an inference posterior or an forecast-revision posterior. 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, the opposite if it is $down$. The updates of forecast revision answers are normalized by $Pr(up|G) - Pr(up|B)$ so that they are comparable to the inference updates. *Rational Update* is the update prescribed by Bayes' rule.

B Additional Analysis on Modes of Behavior

In this section, we provide additional analysis on 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* treatment: subsample with asymmetric priors

Mode	Criterion for answer	*Inference*	*Forecast Revision*
Non-update	= prior	29.9%	23.2%
Exact Representativeness	= 100 if $s_0 > 50$, = 0 if $s_0 < 50$	4.4%	15%
Naive Extrapolation	= s_0	3.5%	9.1%
No Inference-Forecast Gap (excluding the other modes)	inference = forecast revision	2.6%	
Unclassified		60.7%	51.9%
Observations		341	341

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*. Observations with 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. To illustrate, 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 tell 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 participants, we conduct a classification exercise on the participant-part level. Specifically, a participant is classified into a type in a part (*Inference* or *Forecast Revision*) if more than half of her answers in that part are classified into the corresponding mode. Table B3 shows the joint distribution of types across the two parts. The numbers of participants classified in the two parts are 73 and 81, 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	32.1%	25.9%
Exact Representativeness	≈ 100 if $s_0 > 50$, ≈ 0 if $s_0 < 50$	7.2%	20.7%
Naive Extrapolation	$\approx s_0$	3.7%	10.9%
No Inference-Forecast Gap (excluding the other modes)	inference $\approx$ forecast revision	4.5%	
Unclassified		53.0%	40.9%
Observations		1480	1480

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*. Observations with signal equal to 50 are excluded.

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

Forecast Revision type \ *Inference* type	Non-update	Exact Repre-sentativeness	Naive Extrapolation	No Inference-Forecast Gap	Unclassified	Total
Non-update	23	1	1	0	17	41
Exact Representativeness	2	2	0	0	22	26
Naive Extrapolation	4	0	0	0	9	13
No Inference-Forecast Gap	0	0	0	1	0	1
Unclassified	19	0	1	0	101	121
Total	48	3	2	1	149	202

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

Table B4: Modes of behavior in *Cross-variable Forecast*

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*. Observations with signal equal to 50 are excluded.

B.4 Modes of behavior in other treatments

This subsection presents results on the modal behaviors in five treatments: *Cross-variable Forecast*, *Binary Signal*, *Nudge*, *Similarity I*, and *Similarity II*.

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.

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*. Observations with signal equal to 50 are excluded.

C Beliefs without signal realization

In this section, we present results from the parts of our experiment in which participants do not see any signal realization: *Inference Prior*, *Forecast Prior*, and *Expectation Formation*. Figure C1 shows the distribution of answers in *Inference Prior* and *Forecast Prior*. 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*. The distribution of errors is mostly unsystematic.

Like *Forecast Prior*, the experimental part *Expectation Formation* asks about participants' expectations of the outcome without seeing any signal realization. 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 LIE in the *Baseline* treatment and the *Nudge* treatment. The errors are mostly small and unsystematic.

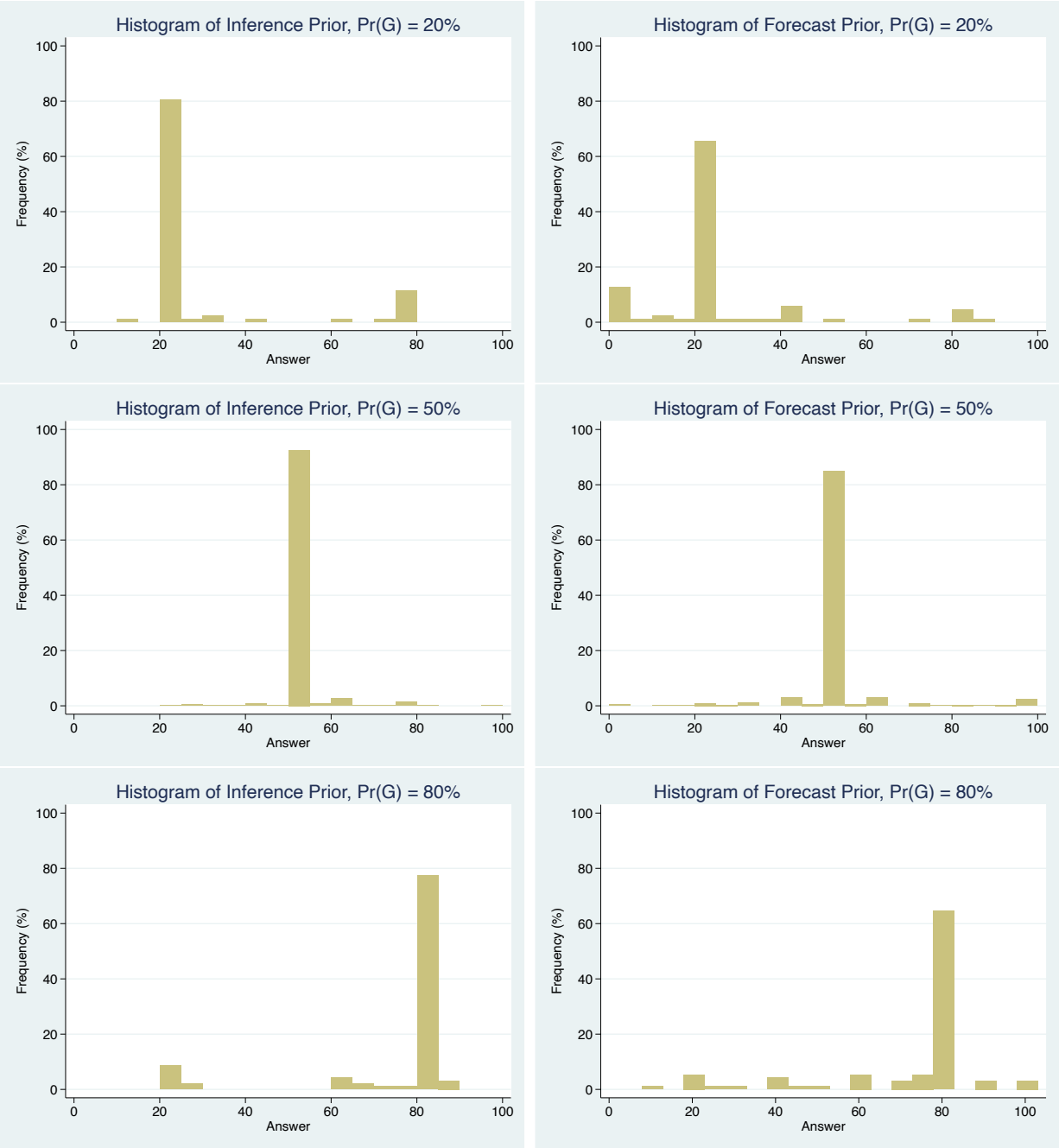

Figure C1: Distributions of answers in *Inference Prior* and *Forecast Prior* in *Baseline*

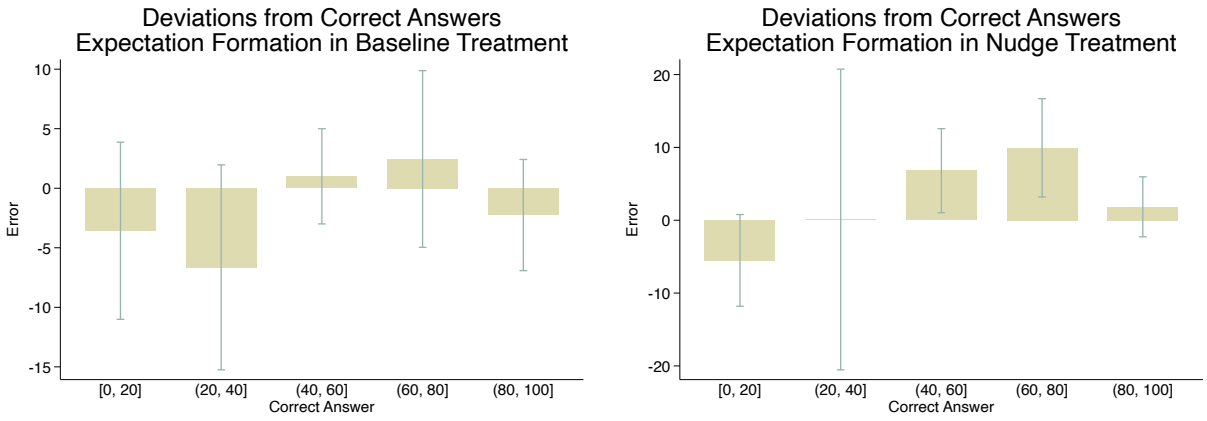

Figure C2: Deviations from LIE in expectation-formation problems

Notes: Standard errors are clustered by participant.