Belief-updating: Inference versus Extrapolation*

Tony Q. Fan[†] Yucheng Liang[‡] Cameron Peng[§]

Preliminary and Incomplete
April 20, 2021

Abstract

Survey forecasts of macroeconomic and financial variables show widespread overreaction to news, but laboratory experiments on belief updating robustly find underinference from signals. We provide new experimental evidence connecting these two seemingly inconsistent phenomena. Building on the classic bookbag-and-poker-chip paradigm, we study how people make inferences *and* revise forecasts in the same information environment. Subjects *underreact* to signals when inferring about fundamentals ("underinference"), but *overreact* to signals when revising forecasts about future outcomes ("overextrapolation"). In the latter task, subjects appear to be using a mix of simplifying heuristics, such as focusing on the representative state (the state most consistent with the signal) and anchoring on the signal. Additional treatments point to distinct mental processes for inference and extrapolation in our environment.

*We thank Nicholas Barberis, Douglas Bernheim, Matthew Gentzkow, Muriel Niederle, and seminar participants at briq, LSE, Stanford, and Yale for helpful comments.

[†]Stanford University.

[‡]briq Institute.

[§]London School of Economics and Political Science.

1 Introduction

Empirical research on belief updating has documented in various settings that reactions to new information typically deviate from Bayes' rule. However, little consensus has been reached on whether the predominant bias is overreaction or underreaction. On the one hand, in surveys, individual forecasts of macroeconomic and financial variables often appear to be overreacting to recent news (e.g., Bordalo et al., 2020). On the other hand, laboratory studies using the bookbag-and-poker-chip paradigm typically find that beliefs about the underlying states underreact to new signals (Benjamin, 2019). This tension in empirical evidence is mirrored in theoretical work. For example, while Kahneman and Tversky (1972) use the representativeness heuristic as the micro-foundation for overreaction, Phillips and Edwards (1966) propose the notion of conservatism to account for underreaction in beliefs.

This paper proposes one way to reconcile these two seemingly contradictory phenomena. We start by pointing out that the two strands of literature entail two related yet distinct types of updating problems (see Figure **??** TBA). In bookbag-and-poker-chip experiments, subjects are asked to update their beliefs about the underlying states. In macroeconomic and financial surveys, forecasters update their expectations about future outcomes. Under perfect Bayesian updating, the two updating problems are equivalent; their only difference, according to most macroeconomic and finance models, is that beliefs about the underlying states are inputs in computing expectations about future outcomes. However, we show, in the same experimental setting, that subjects underreact to signals when updating beliefs about the underlying states but overreact when revising forecasts about future outcomes.

The key idea of our baseline treatment is to compare belief updating about underlying states and about future outcomes in the same environment with a fully specified data-generating process. In each round of the experiment, there is a "firm" with a fixed state which is either good or bad. The firm generates signals which are informative about its state and are framed as its monthly stock price growth; good firms, on average, have higher growth in stock price than bad firms. Subjects do not know the true state, but are given 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. This design follows the bookbag-and-poker-chip paradigm but is framed in economic terms.

The baseline treatment has two main parts: *Inference* and *Extrapolation*. In the *Inference* part, subjects observe one signal realization and then report their updated beliefs about *the state*. In the *Extrapolation* part, subjects also observe one signal realization but instead report their updated expectations about *the next signal*. Given our environment, these two types of beliefs—beliefs about the state and beliefs about the next signal—should be tightly linked: If one believes that the state is good with $p\%$ chance, then by the Law of Iterated Expectations (LoIE), the expectation about the next signal should be $p\% \times 100 + (1 - p\%) \times 0 = p$. We explain this linkage to subjects through experimental instructions and make sure they understand it through comprehension checks.

Despite this tight theoretical connection, there is a significant gap between subjects' beliefs in *Inference* and *Extrapolation*. In *Inference*, 66% of the answers underreact relative to the Bayesian benchmarks while 33% overreact. This systematic underreaction replicates the stylized fact in the bookbag-and-poker-chip literature. By contrast, 46% of the answers in *Extrapolation* underreact while 53% overreact. This *Inference-Extrapolation gap* is validated in a quantitative analysis of subjects' posterior beliefs. Moreover, the gap is robust to i) alternative sample screening criteria, ii) alternative framing, and iii) alternative designs in which the signal follows a binary distribution. We also demonstrate in an additional treatment that, unlike in previous studies in which outcomes and signals are of the same variable, our results are not driven by this outcome-signal similarity.

Interesting patterns emerge when we further examine the distributions of posterior beliefs. In *Inference*, the modal behavior is No Update: in 37 percent of all answers, posterior beliefs are the same as priors. In *Extrapolation*, the fraction of No Update drops to 27 percent. At the same time, two additional modes of behavior stand out. The first mode entails subjects answering the conditional mean of the signal under the state most consistent with the past signal. This, in our setting, is reflected by answering 100 when the signal is better than average and 0 when it is worse than average; 24 percent of all answers fall into this mode. The second mode, constituting 13 percent of all answers, is to report an expectation that equal the past signal. These two modes of behavior are rarely observed in *Inference*, and excluding these answers would completely wipe out the Inference-Extrapolation gap.

We conduct additional analyses to shed light on the underlying mechanisms driving the Inference-Extrapolation gap. The gap should not arise if, in *Extrapolation*, subjects follow a two-step *infer-then-LIE* procedure by i) first updating their beliefs about the states in the same way as in *Inference* and ii) then using these posterior beliefs to *correctly* compute expectations about the next signal under the LIE. The existence of the gap rejects the correct implementation of this two-step procedure. It also leads us to ask whether the gap results from unintentional implementation errors caused by the complexity of the *infer-then-LIE* procedure or the intentional use of alternative procedures for the extrapolation problems. To answer this question, we run an additional treatment where we show subjects their own inference answers when they solve extrapolation problems. This treatment reduces the implementation complexity of the *infer-then-LIE* procedure to the level of a standalone expectation formation problem. Our results show that this treatment does not have any impact on the Inference-Extrapolation gap. Moreover, we check that subjects are largely capable of solving expectation formation problems correctly. Taken together, we conclude that the Inference-Extrapolation gap reflects the intentional use of decision procedures in extrapolation problems that are distinct from the *infer-then-LIE* procedure.

Why don't subjects use the *infer-then-LIE* procedure for extrapolation problems but instead employ distinct heuristics? We hypothesize that it is because the link between *Extrapolation* and *Inference* is not obvious. We test this idea in a final treatment where we make the link obvious to subjects and see if the Inference-Extrapolation gap remains. In this treatment, we tell subjects that a firm's revenue goes up if and only if the firm is good. In the inference questions, we ask subjects to report their beliefs about the states and in the extrapolation questions, we ask them how likely the firm's revenue will go up next month. This treatment makes it obvious that *Extrapolation* and *Inference* are the same problem. The result shows that subjects *underreact* to the same extent in inference and extrapolation problems, making the Inference-Extrapolation gap disappear in this treatment. This result suggests that the inability to conceptually connect the two problems plays a key role in generating the Inference-Extrapolation gap.

This paper contributes to the laboratory literature on belief updating biases (Benjamin, 2019). We replicate the underreaction result from the bookbag-and-poker-chip paradigm and show that, in fact, the same result does not generalize to extrapolation problems asking about future signals. Overreaction in extrapolation problems provides experimental support for overreaction in survey expectations (Greenwood and Shleifer, 2014; Bordalo et al., 2018, 2019, 2020). We further attribute the gap between the two problems to the more complex nature of extrapolation, which induces subjects to use simplifying procedures different from those employed in inference problems. This mechanism also connects to a recent literature on complexity-induced mental models of belief biases (Enke and Zimmermann, 2019; Enke, 2020; Graeber, 2020) and to models of natural expectations (Fuster et al., 2010, 2012).

Our overextrapolation result is reminiscent of the hot hand fallacy (Camerer, 1989; Rabin, 2002; Rabin and Vayanos, 2010). Indeed, the two phenomena may share common psychological underpinnings. One possible explanation for the hot hand fallacy is the gambler's fallacy (and more fundamentally the law of small numbers): The agent *overinfers* about the state from *long streaks* and expects them to continue (Rabin, 2002; Rabin and Vayanos, 2010). This interpretation, however, is inconsistent with both the existence of the Inference-Extrapolation Gap and the fact that overextrapolation occurs after just *one* signal realization in our setting.

Previous experimental work such as Frydman and Nave (2017) and Afrouzi et al. (2020) have also shown systematic overreaction in beliefs. Earlier papers typically use a single time-series as the data generating process (DGP), and the Bayesian benchmark requires a conscious and correct perception of the DGP. In contrast, our environment is static, and the Bayesian benchmark does not require subjects to understand any time-series properties. This simple setting allows for a clean comparison between inference and extrapolation problems. We further show, in an additional treatment, that our results can be generalized to a setting in which signals and outcomes are two different variables. Therefore, the two approaches for studying belief updating concern different information environments and are complementary to each other.

Finally, our results provides experimental support for models of overreaction, including models of diagnostic expectations (Bordalo et al., 2018, 2019, 2020) and of extrapolative beliefs (Barberis et al., 2015, 2018). The two modes of behavior in our extrapolation problems—loading on the more likely state and anchoring on the past signal—directly support the functional forms in models of diagnostic expectations and of mechanical extrapolation, respectively.

The rest of the paper proceeds as follows. Section 2 outlines our experimental design. Section 3 presents the main results. Section 4 discusses the decision procedures used by our subjects, and Section **??** explores possible mechanisms behind our main results. Section 5 discusses the implications of our results, and Section 6 concludes.

2 Experimental Design

2.1 Environment

To compare belief updating in inference and extrapolation problems, we adopt a within-subject experimental design. For each inference problem a subject solves, there is a corresponding extrapolation problem that uses the same information environment with the identical DGP and signal realization.

The baseline treatment has five parts, which are summarized in Table 1, and each part has eight rounds of problems. In each round, subjects are presented with a "firm" randomly drawn from a pool of 20 firms. A firm's state θ is either G(ood) or B(ad). Subjects do not know the state of the drawn firm, but are given the composition of the pool, which specifies the prior distribution over states. Subjects are also given signals s_t, which are informative of the firm's state and framed as the firm's stock price growth in month t. Conditional on the state, signals are i.i.d. across months. Specifically, signals of a good firm follow a normal distribution of $N(100, \sigma^2)$ and signals of a bad firm follow $N(0, \sigma^2)$. Because good firms are more likely to have higher stock price in the future, higher stock price growth is diagnostic of a good firm.[1]

To sum up, the DGP in each round is fully specified by two pieces of information: prior distribution over states $Pr(\theta)$ and conditional distributions of signals $Pr(s|\theta)$. They are presented to subjects graphically using both figures and numbers in one page of display (see Figure 1 for an example). Table 2 summarizes the value of parameters used in each of the eight rounds. Parameter values are kept the same across all five parts except for *Expectation Formation*, which ensures a consistent environment across different tasks; we will elaborate more on *Expectation Formation* below.

In each round of the two main parts, *Inference Posterior* and *Extrapolation Posterior*, subjects first observe the firm's stock price growth in the current month s_0. After seeing the realized signal, subjects report their updated (posterior) beliefs about the state $Pr(\theta|s_0)$ in *Inference Posterior* and their expectations about the firm's stock price growth in the next month $\mathbb{E}(s_1|s_0)$ in *Extrapolation Posterior*. To ensure the comparison between these two parts is apples-to-apples, signal realizations in any two corresponding rounds are set to be the same for each subject.

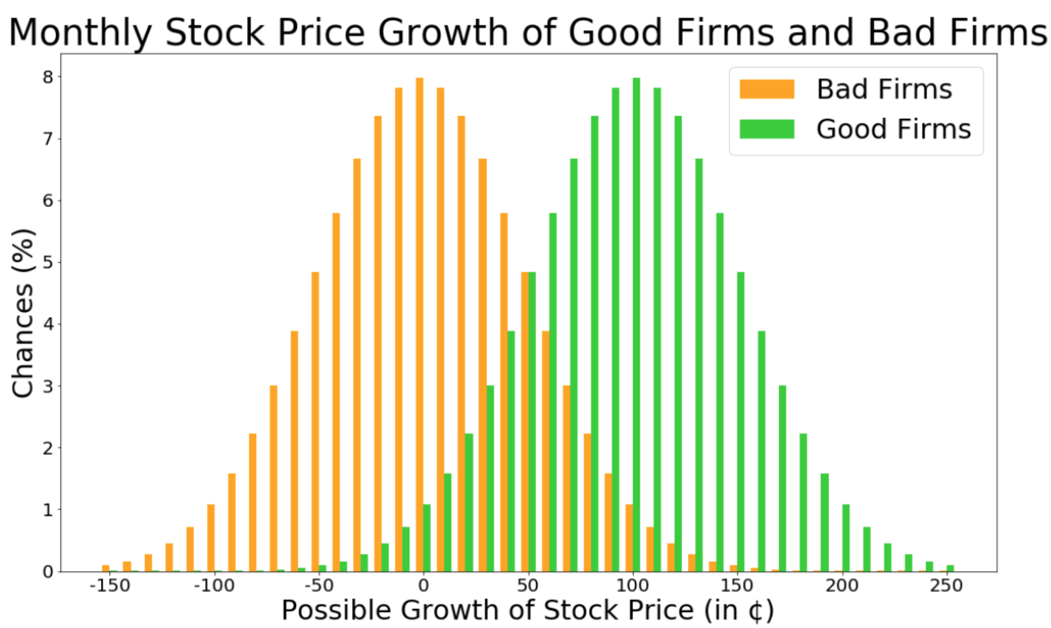
In the other three parts, subjects do not observe any signal realization when beliefs are elicited. Based on the prior distribution of states and the conditional distributions of signals, they report their beliefs about the state $Pr(\theta)$ in *Inference Prior* and expectations about the signal $\mathbb{E}(s_1)$ in *Extrapolation Prior*. These two parts detect errors in the formation of prior beliefs, which are controlled for when measuring belief-updating biases. The last part, *Expectation Formation*, is identical to *Extrapolation Prior*, except for the composition of firms in the pool. Instead of choosing state distribution for subjects as what we do in the other four parts, the composition of firms in this part is determined endogenously by subject's reported posterior beliefs about the states in *Inference Posterior*. For example, if a subject reports a posterior belief of $Pr(\theta = G|s_0) = 40\%$ in a round in *Inference Posterior*, then the pool of firms in the corresponding round in *Expectation Formation* will have $40\% \times 20 = 8$ good firms and 12 bad ones.[2] *Expectation Formation* is de-

[1]In implementing this design, we discretize the supports of normal distributions to multiples of 10 and truncate at both tails. This simplification is further accounted for in our subsequent analysis.

[2]The numbers of good and bad firms in *Expectation Formation* are rounded to the nearest integers if the reported beliefs in *Inference Posterior* are not a multiple of 5%. 14 percent of the answers in *Inference Posterior* are not

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

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



The pool of firms has the following composition.

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

Figure 1: An Example of the Data Generating Process

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

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

signed to understand whether subjects can correctly form expectations about the next signal when the states are distributed according to their inference posteriors.

Subjects need to stay on each page for at least eight seconds before they can type in their answers. This requirement aims to ensure sufficient attention to the problems and prevent click-through behavior. For each subject, we further randomize (i) the order across different DGPs in each part, and (ii) the order across different parts. For the later randomization, we require (i) eliciting priors before eliciting posteriors and (ii) the *Expectation Formation* part comes after the two *Inference* blocks. We focus on three orders of parts —namely, 12345, 12534, and 34125— and, in the Appendix **??**, we show that order effects do not drive our main results. At the end of the experiment, subjects may receive a $5 bonus payment, the chance of which depends on their answer in one randomly selected round through a quadratic rule.[3]

Subjects receive extensive instructions, which explain the details of the tasks and the incentive structure in intuitive terms. In particular, we use several ways to ensure that the properties of the DGPs are understood by subjects. First, we emphasize that the state of a firm is constant across months but the signals are i.i.d. conditional on the state. Second, we also use an example DGP

multiples of 5%, among which half are rounded up and the other half rounded down.

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

Index	$Pr(\theta = G)$	σ
1	50%	5
2	50%	6
3	50%	7
4	50%	8
5	50%	9
6	50%	10
7	80%	10
8	20%	10

Table 2: Parameter values for the data generating process

to illustrate the discretized normal distributions of the signals and to highlight the two different means (0 and 100). Third, we present subjects with two explicit formulae, one for calculating the prior distribution over states from the pool composition ($Pr(\theta = G) = \frac{\text{Number of Good Firms}}{20}$) and one for calculating the expectation about the signal from the belief about the states ($\mathbb{E}(s) = Pr(\theta = G) \times 100$). However, we do not mention or nudge subjects toward any specific belief updating rule. At the end of the instructions, subjects need to answer a set of comprehension questions to test their understanding of the DGPs, the incentive structure, and the two formulae. [MENTION HOT HAND/GAMBLER'S FALLACY QUESTION?] Subjects can only proceed once they have answered all the comprehension questions correctly.

After the five parts, we ask half of our subjects to verbally describe their thought processes in *Inference Posterior* and *Extrapolation Posterior*. Finally, the experiment ends with a questionnaire on demographic information.

Building off of the baseline treatment, we implement a few straightforward extensions and robustness checks. First, we ran versions of the baseline treatment with the signals framed as monthly revenue growth instead of stock price growth. Second, we ran versions of the baseline treatment in which questions about posterior beliefs are framed as expectations of the *last* signal s_{-1} ("stock price/revenue growth in the previous month") instead of the *next* signal s_1. In Online Appendix **??**, we show that the results are similar in all these extensions; thus, we pool the data from all versions of the baseline treatment for our main results.

2.2 Theoretical benchmarks

Now we define the theoretical benchmarks in our experiment. In an inference problem, a rational decision maker updates her belief about the states by applying Bayes' rule to the DGP and the realized signal s_0:

$$Pr(\theta = G|s_0) = \frac{Pr(\theta = G) \cdot Pr(s_0|\theta = G)}{Pr(\theta = G) \cdot Pr(s_0|\theta = G) + Pr(\theta = B) \cdot Pr(s_0|\theta = B)} \tag{1}$$

For the extrapolation problem, the rational answer can be derived by first calculating the Bayesian posterior over the states and then apply the Law of Iterated Expectations (LoIE) to obtain the expectation of the next signal. This procedure leads to the following equation.

$$\mathbb{E}(s_1|s_0) = Pr(G|s_0) \times \mathbb{E}(s_1|\theta = G) + Pr(B|s_0) \times \mathbb{E}(s_1|\theta = B) = \underbrace{\underbrace{Pr(G|s_0)}_{\text{Step 1}} \times 100}_{\text{Step 2}} \tag{2}$$

Equations (1) and (2) define the **rational benchmarks** for our inference and extrapolation problems, respectively. Decades of empirical research have shown that belief updating often deviates from the rational benchmarks.[4] To study such deviations, we further consider two classifications. A posterior belief is considered *overreaction* if it updates too much based on the signal; that is, the revision of belief goes in the same direction as the rational benchmark but by an amount that is too

[4] CITATIONS TBA

much:

$$(\text{Posterior Belief} - \text{Rational benchmark}) \times (\text{Rational benchmark} - \text{Prior}) > 0. \tag{3}$$

We define *underreaction* with a similar condition.

$$(\text{Posterior Belief} - \text{Rational benchmark}) \times (\text{Rational benchmark} - \text{Prior}) < 0.^{5} \tag{4}$$

Given our motivation of comparing between inference and extrapolation, another metric of interest is the gap in posterior beliefs between the two problems. Denote the *actual* posterior beliefs in the inference and extrapolation problems by $Pr^A(\theta|s_0)$ and $\mathbb{E}^A(s_1|s_0)$. Then, we define the **no Inference-Extrapolation gap benchmark** as the following condition.

$$\mathbb{E}^A(s_1|s_0) = Pr^A(\theta = G|s_0) \times 100. \tag{5}$$

Under this condition, a decision maker who underreacts in an inference problem will underreact by exactly the same amount in the corresponding extrapolation problem.

The no Inference-Extrapolation gap condition is satisfied if a decision maker uses the following *infer-then-LoIE* procedure to solve the extrapolation problems.

- Step 1 (infer): Calculate the posterior over the states using the same (and possibly non-Bayesian) updating rule as in the inference problem.
- Step 2 (LoIE): Apply the LoIE to the Step 1 output to obtain the expectation over the next signal.

Nevertheless, an Inference-Extrapolation gap can arise for several reasons. For example, it could be that the LoIE is not applied correctly in Step 2 of extrapolation problems. Alternatively, the updating rule used in Step 1 of extrapolation may be different from that in inference problems. Moreover, it could be that people do not use the Inference-Expectation two-step procedure

[5] TBA: explain base-rate neglect, wrong direction, out-of-bound, not exhaustive (signal = 50).

at all when they solve extrapolation problems. Under these possibilities, a decision maker can simultaneously exhibit distinct biases in inference and extrapolation problems.

2.3 Procedural details

We programmed our experiment using oTree (Chen et al., 2016). For the baseline treatment, we recruited 202 subjects through Prolific, an online platform designed for social science research.[6] 120 of them saw signals framed as monthly revenue growth and 82 saw signals framed as stock price growth. For 40 subjects, questions in Parts E, E0, and E1 asked about expectations of the *last* signal ("stock price/revenue growth in the previous month") instead of the *next* signal. Seventy-two subjects went through five parts of the experiment in the order of 12345, 73 subjects went in the order of 12534, and 57 went in the order of 34125. Subjects spent on average about 30 minutes on the experiment and earned an average payment of \$7.15 including a \$5 base payment.

3 Main Results

In this section, we present the results from our baseline treatment with a focus on the Inference-Extrapolation gap.

3.1 The Inference-Extrapolation gap

3.1.1 Evidence from classification

Based on Equations (3) and (4), we classify all the answers in *Inference Posterior* and *Extrapolation Posterior* into three types: *Underreact*, *Exactly Bayesian*, and *Overreact*. For each answer the subject gives, we compare it to the rational benchmark based on the subject's own prior elicited in the corresponding prior part.[7] In other words, the rational benchmark is taken as the Bayesian posterior implied by the subjective prior rather than the objective prior.

[6] See Palan and Schitter (2018) for using Prolific as a subject pool. We only recruited US subjects who have completed more than 100 tasks on Prolific and have an approval rate of at least 99%.

[7] Answers in which subjects do not update at all are considered *Underreact*.

Figure 2 shows the overall classification results. Results from the inference problems replicate findings from the classic "bookbag-and-poker-chip" literature: subjects overwhelmingly underreact to new information and update too little about the firm's underlying state. Out of all the answers, 67% imply underreaction while 33% imply overreaction; virtually no answers are considered *Exactly Bayesian*. These patterns, however, flip in the extrapolation problems: There, subjects appear to overreact to new information systematically. Around 53% of the answers indicate overreaction to the signal while 46% indicate underreaction. The Inference-Extrapolation gap of 20% in the rate of overreaction is highly statistically significant ($p < 0.01$).

To further control for the potential influence of base rate neglect, we repeat the same exercise using two subsamples. The first subsample includes problems in which the *objective* priors for the two states are 50%-50%, and the second subsample is further restricted to subject-problem pairs in which the subjective priors provided by the subject are also 50%-50% (in both *Inference Prior* and *Extrapolation Prior*). With a 50%-50% prior, base rate neglect is irrelevant for belief updating, which allows for a cleaner detection of updating biases. In both subsamples, we observe a similar pattern: subjects on average underreact in *Inference Posterior* and overreact in *Extrapolation Posterior*.

3.1.2 Evidence from posterior beliefs

We provide additional evidence on the Inference-Extrapolation gap by studying subjects' posterior beliefs quantitatively. We examine posterior beliefs about the probability of the state most consistent with the signal (referred to as the "signal-consistent" state henceforth): Because of symmetry around the average signal of 50 cents, the signal-consistent state is effectively the Good state when the signal is higher than 50 cents and the Bad state when it is lower than 50 cents. In *Inference Posterior*, subjects directly report their posterior beliefs about the firm's state. In *Extrapolation Posterior*, subjects report their forecasts about the firm's future signal, and we compute the posterior beliefs about the states *implied by* their forecasts.

Figure 3 shows the average posterior beliefs about the signal-consistent state, along with the average prior beliefs and the average implied Bayesian posteriors, separately for the inference and the extrapolation problems. In both types of problems, subjects start with an average prior of around 50.5% about the signal-consistent state[8] and, under Bayesian updating, the average posteriors should be around 73-74%. However, the actual posteriors are drastically different: they average to 66% in *Inference Posterior*, indicating underreaction, and to 81% in *Extrapolation Posterior*, suggesting overreaction. This gap of 15% in the average posteriors is highly statistically significant ($p < 0.01$).

3.2 Heterogeneity of the gap

In this section, we show that the Inference-Extrapolation gap is a robust phenomenon in various cuts of the data. We start by separately examining each of the eight problems with different parameters. The results are plotted in Panel A and Panel B of Figure 4. While the eight problems differ in the prior distribution over states and the distribution of signals, we find that the gap is robust across the board: For all problems, subjects are more likely to overreact in *Extrapolation Posterior* and underreact in *Inference Posterior*.

For the subsample with both objective and subjective priors of 50%-50%, we further examine how the Inference-Extrapolation gap depends on the strength or diagnosticity of the signal. We measure signal strength by the implied Bayesian posterior of the signal-consistent state after seeing the signal: The higher it is, the more information the signal conveys about the underlying state. Panel C of Figure 4 plots the results. Overall, there is a larger Inference-Extrapolation gap when the signal is more diagnostic. There is no gap for the weakest signals, as there is less underinference (Benjamin, 2019) and no systematic overextrapolation for those signals.

We also vary the order in which the three blocks are presented, and the gap is robust to potential order effects. As Panel D of Figure 4 shows, there is a large and statistically significant gap for all the three possible orders the different parts are presented.

[8] This is only slightly higher than 50% because only one fourth of all problems in our experiment do not have a prior of 50%-50%, and even in those problems, many signals are consistent with the a priori unlikely state given the inherent noise in the signals.

3.3 Distribution of posterior beliefs

In this section, we examine the distributions of posterior beliefs to explore possible heuristics driving the Inference-Extrapolation gap. For the subsample with objective priors of 50%-50%, Figure 5 shows the distributions of posterior beliefs in *Inference Posterior* and *Extrapolation Posterior*, separately for "good" signals (signals higher than 50 cents) and "bad" signals (signals lower than 50 cents).

These plots reveal two patterns. First, subjects are more likely to not update their beliefs at all in *Inference Posterior* than in *Extrapolation Posterior*: upon receiving a signal—either good or bad—there is an around 30% chance that a subject does not update her prior in *Inference Posterior*. The chance of no updating, however, is around 25% in *Extrapolation Posterior*, consistent with subjects reacting more strongly to the signal in the latter case.

Second, a substantial fraction of the forecasts in *Extrapolation Posterior* imply a posterior belief of 100% (or higher) for the signal-consistent state, which is a strong driving force of the Inference-Extrapolation gap. Intuitively, bunching exactly at 100% suggests that when a subject sees a good (bad) signal diagnostic of the Good (Bad) state, she revises her forecast about the next signal as if the state is Good (Bad) for sure and uses the conditional mean of signals in the Good (Bad) state as her forecast. We term this behavior "Loading on the Signal-consistent State."

Another notable pattern among answers in *Extrapolation Posterior* is that some subjects report exactly the past signal as their forecast about the next signal. We call this behavior "Anchoring on the Signal". These subjects likely adopt a simplifying heuristic or expect the signal to be extremely persistent.

In Table 3.3, we formally define *types* of answers consistent with the modes of behavior discussed above to study their prevalence. For example, an answer is consistent with "Loading on the Signal-consistent State" if it equals the expected signal for a good (bad) firm and the observed signal is better (worse) than average; and an answer is consistent with "Anchoring on the Signal" if it equals the signal. Table 6 shows the breakdown among all the answers into different types. Both the "Loading on the Signal-consistent State" type and the "Anchoring on the Sig-

Table 3: Definition of types.

Type	Extrapolation $E[s_1 \mid s_0]$	Inference $Pr(\text{Good} \mid s_0)$
Loading on the Signal-consistent State	100 if s_0 > 50, 0 if s_0 < 50	100% if s_0 > 50, 0% if s_0 < 50
Anchoring on the Signal	s_0	$s_0\%$
No Update	Same as Prior	
No Inference-Extrapolation Gap (excluding No Update)	$E[s_1 \mid s_0] = Pr(\text{Good} \mid s_0) \times 100$ and not Same as Prior	

nal" type constitute a significant portion of the answers in *Extrapolation Posterior*, with "Loading on the Signal-consistent State" being more common. In comparison, these types are virtually non-existent in *Inference Posterior*; therefore, more answers are not classified into any type in *Inference Posterior* than in *Extrapolation Posterior*.

We also conduct a subject-level classification exercise, in which we classify a *subject* into a part-specific type if more than half of her answers in that part fall into that type. Table 6 shows the results. Around 13% of the subjects are classified as the "Loading on the Signal-consistent State" type, while around 6% of them fall into the "Anchoring on the Signal" type.

3.4 Robustness

We now discuss the robustness of the Inference-Extrapolation gap. First of all, in our baseline treatment, the gap is qualitatively robust to various sample screening criteria: dropping all problems in which either the inference posterior or the extrapolation posterior indicates updating in the wrong direction and all problems in which either the extrapolation prior or the extrapolation posterior falls outside the [0, 100] range; and restricting the sample to subject-problem pairs where the subject gives objectively correct answers for both the inference prior and the extrapolation prior.

In different versions of the baseline treatment, we show that the gap is robust to higher monetary incentives and changes in the economic environment. In one version, we double the size of bonus payments so that subjects potentially have even stronger incentives to give correct answers, and we find similar results. In a second version, we use the firm's stock return (instead of stock price growth) as the signal, and again find that subjects effectively exhibit the same gap in their responses. In a third version, we frame the signal as the firm's revenue growth (instead of stock price growth), and subjects do not appear to respond to this change in framing. In a fourth version, we ask about expectations of the *last* signal instead of the next signal in the extrapolation problems, which does not change the theoretical benchmarks; results are still virtually unchanged.

In a final version, we change the signal distribution from a continuous distribution to a binary distribution; this version more closely aligns with the setup of the classic "bookbag-and-poker-chip" paradigm. More detailed results are reported in the Online Appendix. In a nutshell, we find a similar gap between Inference and Extrapolation, albeit with a smaller magnitude.

3.5 Signal-Outcome Similarity [STREAMLINE WITH PREVIOUS PARAGRAGHS TBA]

One potential explanation for the gap is the similarity between the signal and the outcome to be predicted (Kahneman and Tversky, 1972). According to this explanation, subjects overreact to the signal in *Extrapolation Posterior* because the same variable is also the outcome to be predicted in the task; this similarity may lead subjects to perceive the signal as more informative. In comparison, in *Inference Posterior*, subjects are asked to report their beliefs about the states, which are a different variable from the signal.

To test this hypothesis, we run an additional treatment, the *Dissimilar Outcome* treatment, with 100 subjects. In this treatment, we change the outcome to be predicted in the extrapolation problems from the next signal to a third variable that is determined by the state. Specifically, we frame the outcome as the firm's revenue growth when the signal is the firm's stock price growth, and vice versa. Moreover, the outcome has 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 it is similar to the state in terms of distribution.

Figure 7 shows the results from the *Dissimilar Outcome* treatment. Similar to the baseline treatment, there is a statistically significant Inference-Extrapolation gap whose magnitude is even larger: For example, only 29% of the inference posteriors imply overreaction while 57% of the extrapolation posteriors do ($p < 0.01$). This suggests that signal-outcome similarity does not explain our main finding.

Besides ruling out a potential mechanism driving the Inference-Extrapolation gap, the *Dissimilar Outcome* treatment also broadens the external relevance of our results. In most empirical settings, the forecasters' information sets are not limited to past observations of the variables to be predicted. The results in the *Dissimilar Outcome* treatment suggest that the Inference-Extrapolation gap can be an explanation for overreactions in these settings as well (e.g. Bordalo et al., 2020).[9]

4 Decision Procedures

4.1 Distinct heuristics or implementation errors?

As mentioned in Section 2.2, the Inference-Extrapolation gap should not arise if subjects, in answering an extrapolation question, first update their beliefs about the states as in the inference problem and then correctly apply the LoIE to form expectations about the next signal. The evidence we have documented so far on the Inference-Extrapolation gap clearly rejects this *infer-then-LoIE* procedure. Moreover, the prevalence of behaviors such as Focusing on the Signal-Consistent State and Anchoring on the Signal in *Extrapolation* hints at the use of distinct decision procedures. However, before studying this latter possibility, we first examine imperfect implementations of the *infer-then-LoIE* procedure more closely to more carefully rule it out. It could be the case that subjects intend to follow the *infer-then-LoIE* procedure, but because of its complexity, they make implementation errors which lead to the Inference-Extrapolation gap. For instance, a decision-maker may be capable of forming probabilistic beliefs about the states when the inference question is all that is asked. But when implementing the two-step *infer-then-LoIE* procedure for the extrapola-

[9] It is unclear whether previous results on univariate time-series forecasts are applicable to these settings.

tion problem, she may only have enough cognitive bandwidth to form binary beliefs ("The firm is good" or "The firm is bad") for the first step. This error can lead to behaviors that look like Focusing on the Signal-Consistent State.

We run an additional treatment, *Nudge* (N=99), to test the hypothesis that the Inference-Extrapolation gap results from complexity-induced errors when subjects try to implement the *infer-then-LoIE* procedure. In each round of this treatment, subjects answer two questions—they report their beliefs about the states first and then their expectations about the next signal.[10] For subjects who intend to follow the *infer-then-LoIE* procedure, having the *Inference* answer readily available when solving the *Extrapolation* problem makes implementing the two-step procedure no more complex than solving a standalone expectation formation problem. Indeed, with the answer from inference, completing the *infer-then-LoIE* procedure only takes multiplication by 100. This reduction in complexity should mitigate the hypothesized implementation errors and reduce the Inference-Extrapolation gap.

Table **??** shows the results from the *Nudge* treatment. Answers to the inference questions are very similar to those in the *Baseline* treatment. The proportions of underreaction and overreaction are virtually unchanged. For extrapolation problems, displaying inference answers on the same page has very little effects on the answers. Only a small proportion of subjects give answers that exhibit no Inference-Extrapolation gap. These subjects make virtually no difference to the aggregate-level Inference-Extrapolation gap.

How can one explain the ineffectiveness of the *Nudge* treatment? One possibility is that while the treatment indeed makes the *infer-then-LoIE* procedure no more complex than solving a standalone expectation formation problem, even the latter is too complex and error-prone for subjects. To test this possibility, in another part of the *Nudge* treatment called *Expectation Formation*, we ask subjects to report their expectations of the next signal without showing them any signal realization. Moreover, for each subject, we endogenously set the distribution over states in an expectation

[10]Subjects have to stay on the page for 8 seconds before answering each question. Once a subject submits her answer to the first question, she can still see her answer but cannot revise it. The second question only appears after the answer to the first question is submitted.

formation problem to match the posterior belief the subject reported in the corresponding inference problem. For example, if a subject 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-Extrapolation gap in *Nudge* can be attributed to mistakes in expectation formation.

Figure **??** shows that the average deviations from LoIE in expectation formation problems are small. Moreover, comparing the average beliefs in inference, extrapolation, and expectation formation problems, mistakes in expectation formation can account for only 19% of the Inference-Extrapolation gap. These results indicate that mistakes in standalone expectation formation problems do not explain the null effect of the *Nudge* treatment on the Inference-Extrapolation gap.

Taken together, results from the *Nudge* treatment reject the hypothesis that the Inference-Extrapolation gap stems from unintentional errors when subjects try to implement the *infer-then-LoIE* decision procedure in extrapolation problems. Rather, the gap is likely a result of deliberate use of alternative procedures that are distinct from the *infer-then-LoIE* procedure.

4.2 Evidence from subjective responses

[TBA] The data pattern is also validated by the reflections subjects write about their reasoning processes. [TBA]

4.3 Obviousness of the Connection between Inference and Extrapolation

Why do subjects use heuristics for extrapolation problems that seem detached from their inference answers? One hypothesis is that they do not recognize the connection between the two kinds of problems. To investigate this hypothesis, we implement an additional treatment, *Obvious Connection* ($N = 30$). This treatment is identical to *Dissimilar Outcome* except that in Parts E0, E and E1, we tell subjects that the predicted outcome is positive if and only if the state is G, and we ask for the posterior probability that the next outcome is positive. As the name of the treatment suggests, we design the treatment to make it obvious that the inference and extrapolation questions are asking about the same event.

Figure **??** shows the results in *Obvious Connection*. With the predicted outcome obviously connected to the state, the Inference-Extrapolation gap completely vanishes, and we obtain the familiar underreaction pattern in the extrapolation problems. This result confirms the hypothesis that subjects use distinctive heuristics to solve inference and extrapolation problems because they fail to recognize their conceptual connection. One caveat of the result is that while it provides an explanation for why the decision procedures for extrapolation problems appear disconnected from answers to the inference questions, it does not explain why these decision procedures happen to be overreacting.

4.4 Other Explanations [MOVE TO SECTION 3 TBA]

The *Inference-Extrapolation* gap is robust to several other considerations. First, it remains when the variable of interest changes from stock price change to firm revenue. Similarly, it remains when stock price change is replaced by stock return. This suggests that the gap it not domain-specific and applies more generally to other domains.

[Do we have any other treatments/robustness checks?]

5 Discussion

In this section, we discuss the implications and limitations of our results. Sections 5.2 to 5.4 discuss possible implications for four strands of the literature: experimental evidence on belief-updating, models of investor beliefs, models of Bayesian learning, and evidence from survey expectations. Section 5.5 discusses the limitations of our experimental setting.

5.1 Experimental Evidence on Beliefs

The experimental evidence from the “bookbag-and-poker-chip” paradigm, with few exceptions (citation), shows that people systematically underreact to new information in well-defined

Bayesian benchmark. Other psychologists such as Kahneman and Tversky have questioned this result, citing that it is inconsistent with the daily observation that people often jump to conclusions too quickly. This motivated their influential work on "representativeness" which has been used to provide micro-foundation for studying overreaction in macroeconomic and finance models. We show that the two paradigms—"bookbag-and-poker-chip" and "representativeness"—can, in fact, be consistent with each other. By changing the nature of the task from making inference to forming expectations in the "bookbag-and-poker-chip" paradigm, we show that subjects overreact to signals when forming expectations, in a well-defined Bayesian benchmark. Therefore, we provide experimental support for models of overreaction based on, for example, diagnostic expectations and extrapolative beliefs; we discuss this in greater detail below.

Task-dependent reaction to signals further suggest that the way people respond to new information fundamentally depends on the nature of the task. In particular, when the task becomes more complex—as in "Extrapolation" of our experiment—people start to mentally simplify the task and employ heuristics that lead to overreaction. This observation echoes the literature of natural expectations, according to which the mental simplification of a complex process may lead to overreaction (Fuster et al. (2010, 2012)). The law of small numbers literature has a similar intuition: when people know less about the underlying data-generating process, they put too much emphasis on the signals, leading to over-extrapolation of past trends (Rabin (2002); Rabin and Vayanos (2010)).

5.2 Models of Overreaction

Our experiment not only supports models of overreaction, but also provides evidence suggestive of specific functional forms that can be used to describe the belief-formation process. In particular, we find support for two types of belief-formation processes—mechanical extrapolation and diagnostic expectations—which happen to be among the most popular frameworks for modeling investor beliefs. Mechanical extrapolation suggests that people naively use past outcomes to forecast outcomes, and this is consistent with "anchoring on the signal." Diagnostic expectations suggest that people overly focus on the state of which the signal is more diagnostic—in our setting, this is the state that is consistent with the signal—and this is consistent with "loading on the more likely state." While both types of patterns are present, "loading on the more likely state" appears to be more commonly used by the subjects.

While our evidence is consistent with diagnostic expectations, the interpretation is fundamentally different. The motivation for diagnostic expectations is overinference: upon receiving a signal, an individual infers too much from it in figuring out the underlying states. In other words, in Equation 2, errors occur in Step 1, not in Step 2. Our evidence, however, suggests the failure of two-step reasoning; in fact, if two-step reasoning is work, the fact that subjects under-infer in Step 1 suggests that they should overreact in *Extrapolation* as well. Therefore, although our evidence supports diagnostic expectations as a modeling choice, it casts doubt on representativeness as the underlying driving force.

5.3 Violation of the Law of Iterated Expectations

5.4 Survey Expectations

Recent literature documents systematic deviation from the "full-information-rational-expectation" (FIRE) benchmark in survey forecasts of macroeconomic and financial variables. While there is some evidence of underreaction (e.g., Coibion and Gorodnichenko (2015)), the overall evidence—especially for financial variables such as stock and bond returns—points to overreaction as the more dominant theme. In this regard, our paper joins earlier experimental studies such as Afrouzi et al. (2020) in providing experimental evidence for overreaction. As discussed, a key difference between our setting and those in earlier studies is that ours is not time-series and has a well-defined Bayesian benchmark that has been commonly used in experimental studies for decades.

5.5 Limitations

While our experimental evidence is consistent with overreaction in survey expectations, it does not mean that survey takers necessarily use the same heuristics we discover in making their forecasts. The two settings are different in many other aspects, all of which could be driving overreaction in the field. The data-generating process, for example, is much more complex in reality than in our simple experimental setting. The subjects are also of a different pool, as many surveys are done by professional forecasters or investors who possess a good understanding about the financial market.

We believe our evidence can still speak to—at least partially—what is going on in the field for the following reasons. First, while survey takers are more sophisticated, most market participants are households and closer to the subjects we study. It is also worth noting that survey-based expectations often exhibit high correlation between retails and institutions. Second, with a more complex data-generating process in reality, the mechanism we propose in the paper could be all the more important. If people can't deal with these complex problems mentally, they tend to resort to simplifying heuristics like the ones we discovered in this paper. Even when a small fraction of the population engages in such behavior, they can collectively bias the forecasts in significant ways.

6 Conclusion

References

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

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.

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

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

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

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

C. F. Camerer. Does the basketball market believe in the 'hot hand,'? *American Economic Review*, 79(5):1257–1261, 1989.

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.

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

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

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

A. Fuster, D. Laibson, and B. Mendel. Natural expectations and macroeconomic fluctuations. *Journal of Economic Perspectives*, 24(4):67–84, 2010.

A. Fuster, B. Hebert, and D. Laibson. Natural expectations, macroeconomic dynamics, and asset pricing. *NBER Macroeconomics Annual*, 26(1):1–48, 2012.

T. Graeber. Inattentive inference. *Available at SSRN 3658112*, 2020.

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

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

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

S. Palan and C. Schitter. Prolific.ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018. ISSN 2214-6350. doi: https://doi.org/10.1016/j.jbef.2017.12.004. URL `https://www.sciencedirect.com/science/article/pii/S2214635017300989`.

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.

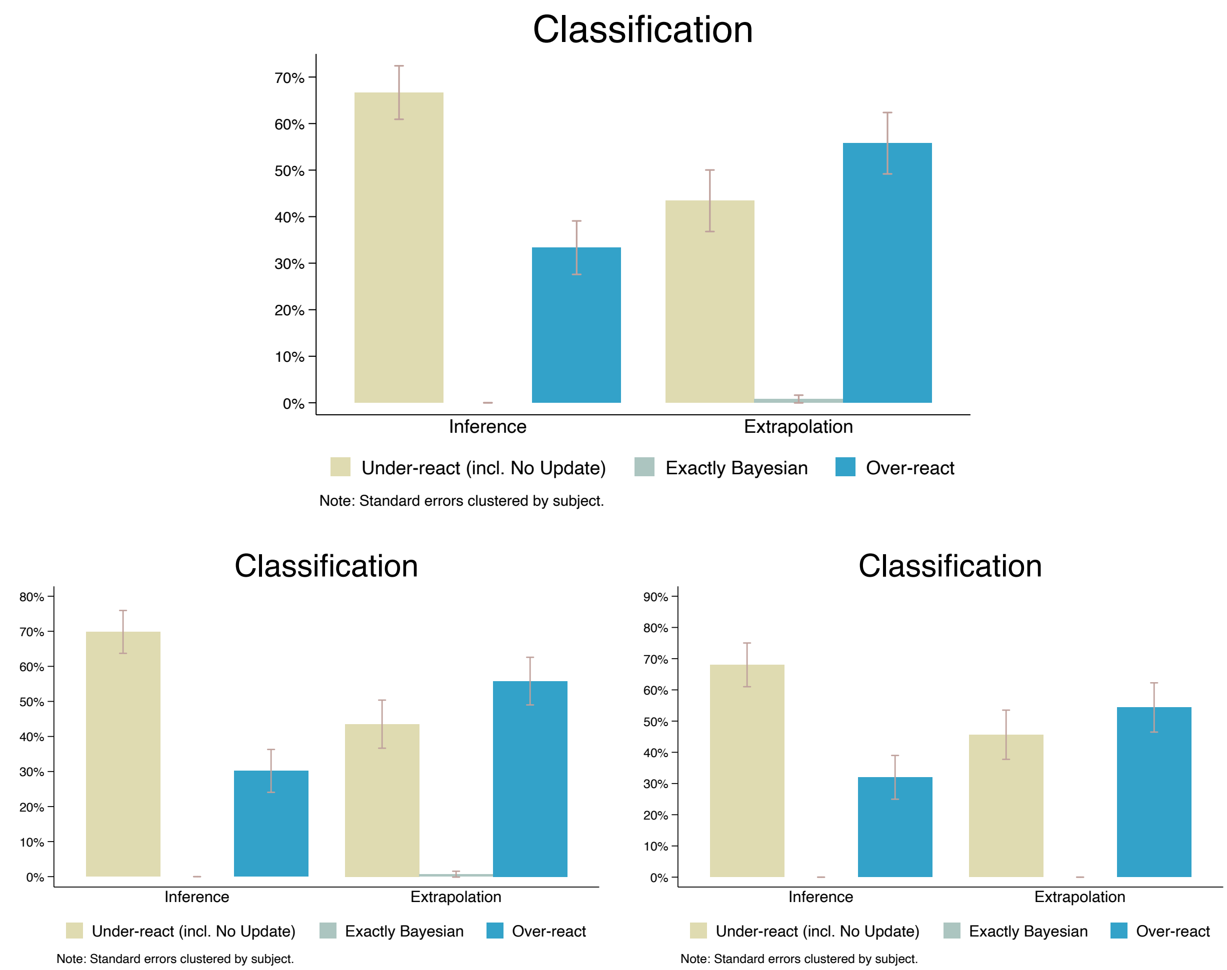


Figure 2: Classification of responses, full sample.

Figures and Tables

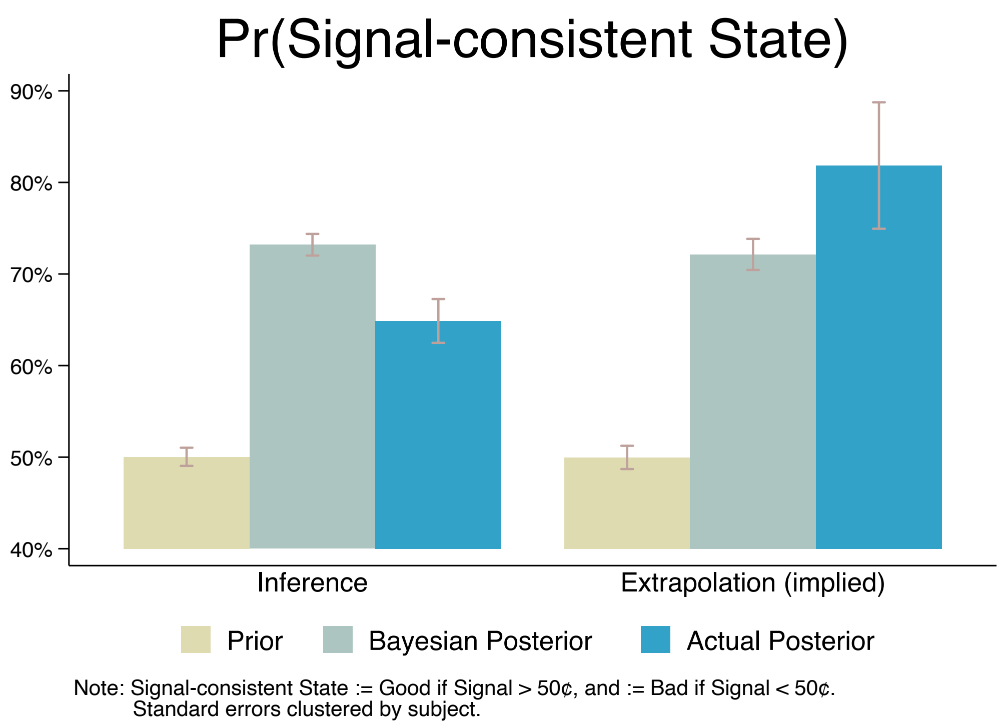


Figure 3: Distribution of posterior beliefs.

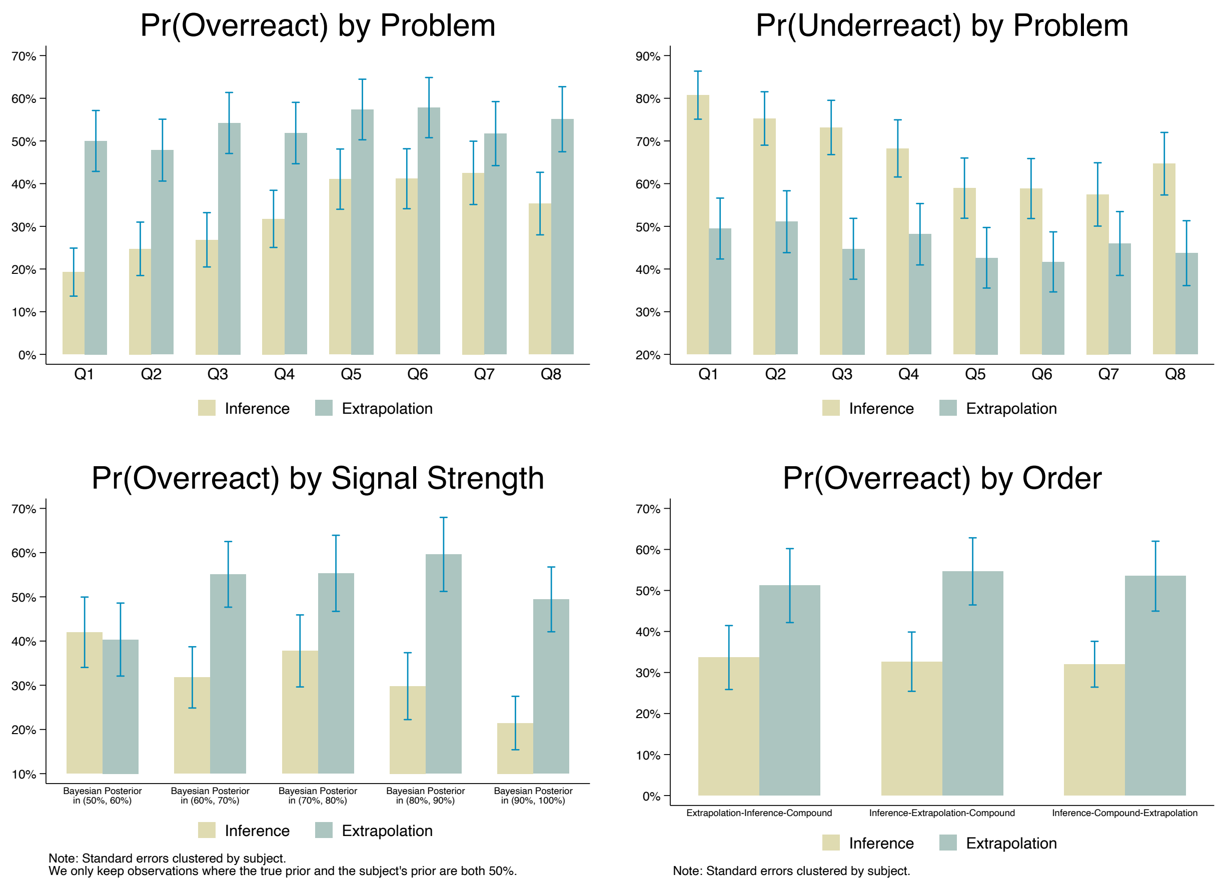


Figure 4: Heterogeneity of the Inference-Extrapolation gap.

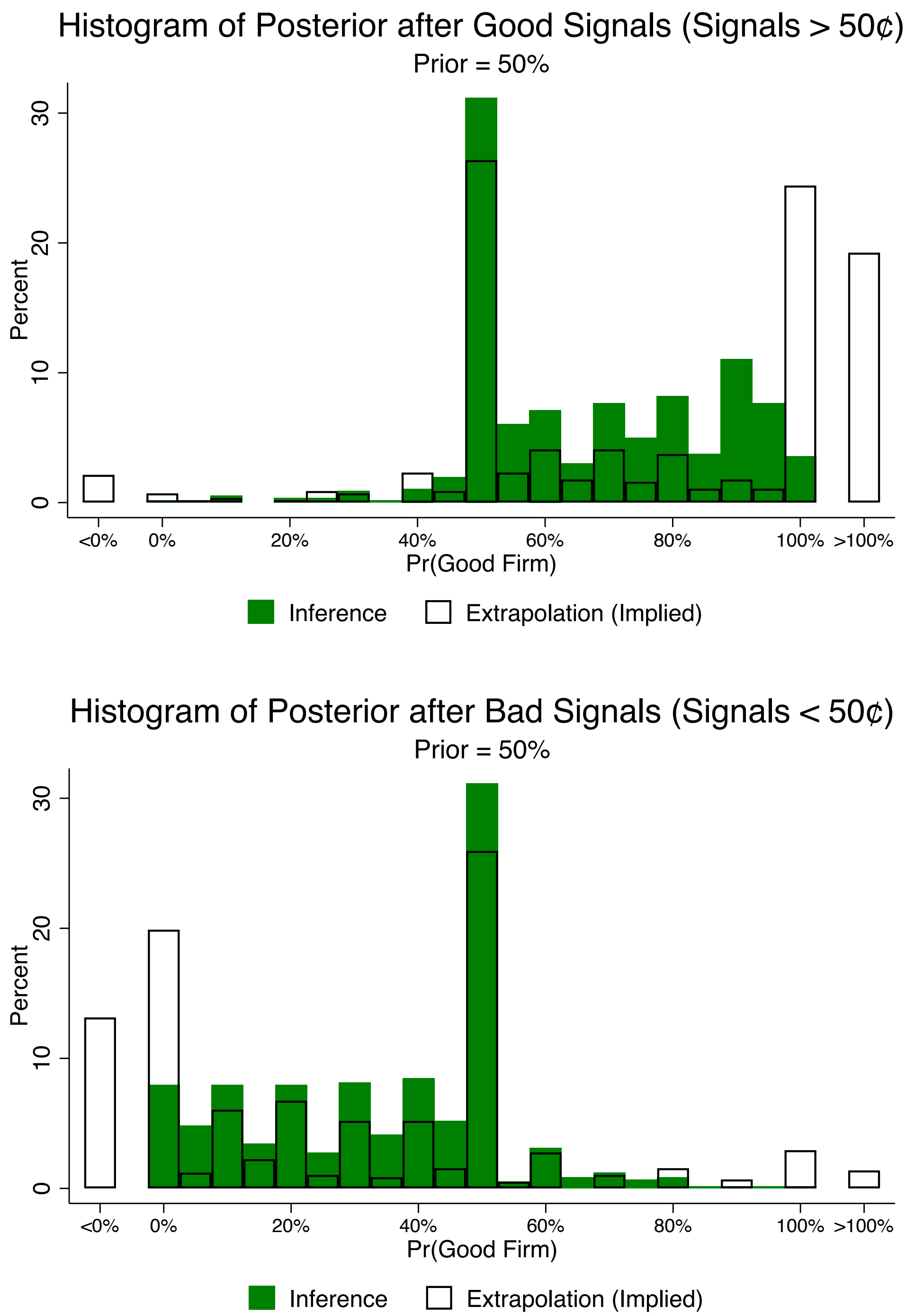
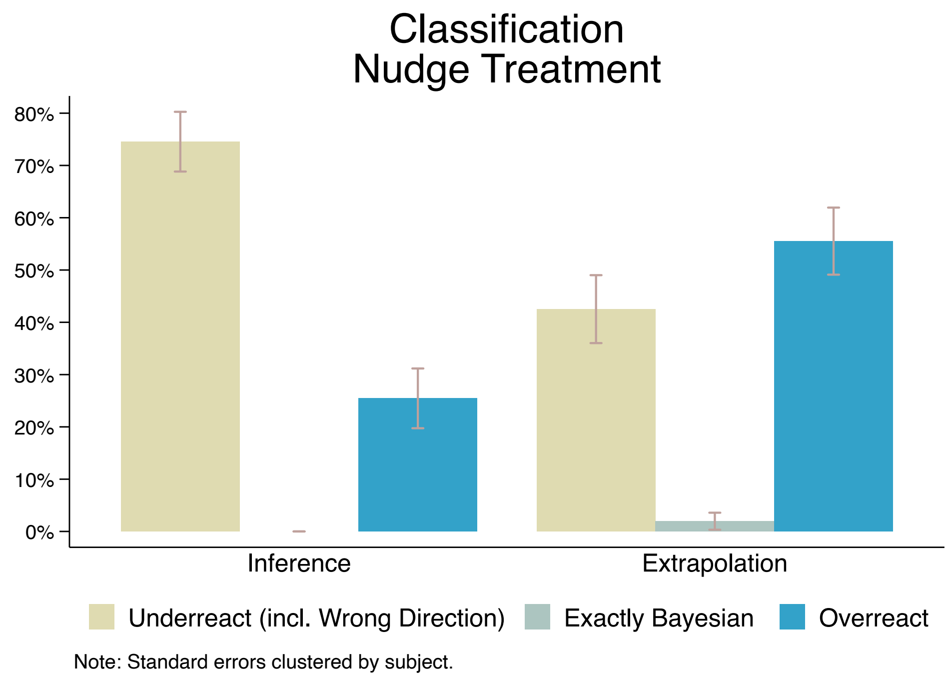


Figure 5: Distribution of posteriors.



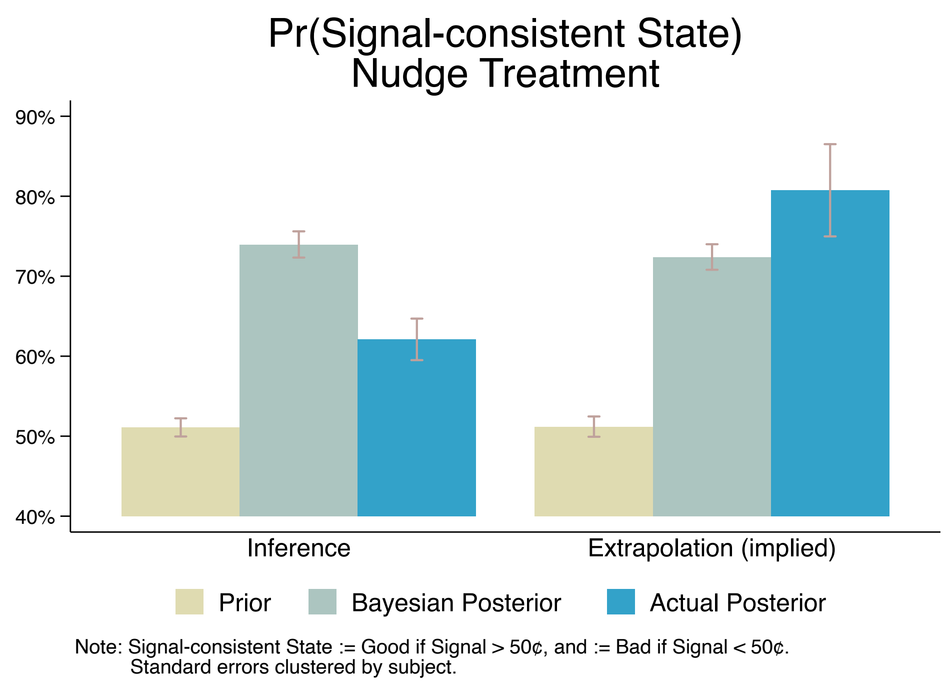


Figure 6: Classification and distribution of posteriors in the *Nudge* treatment.

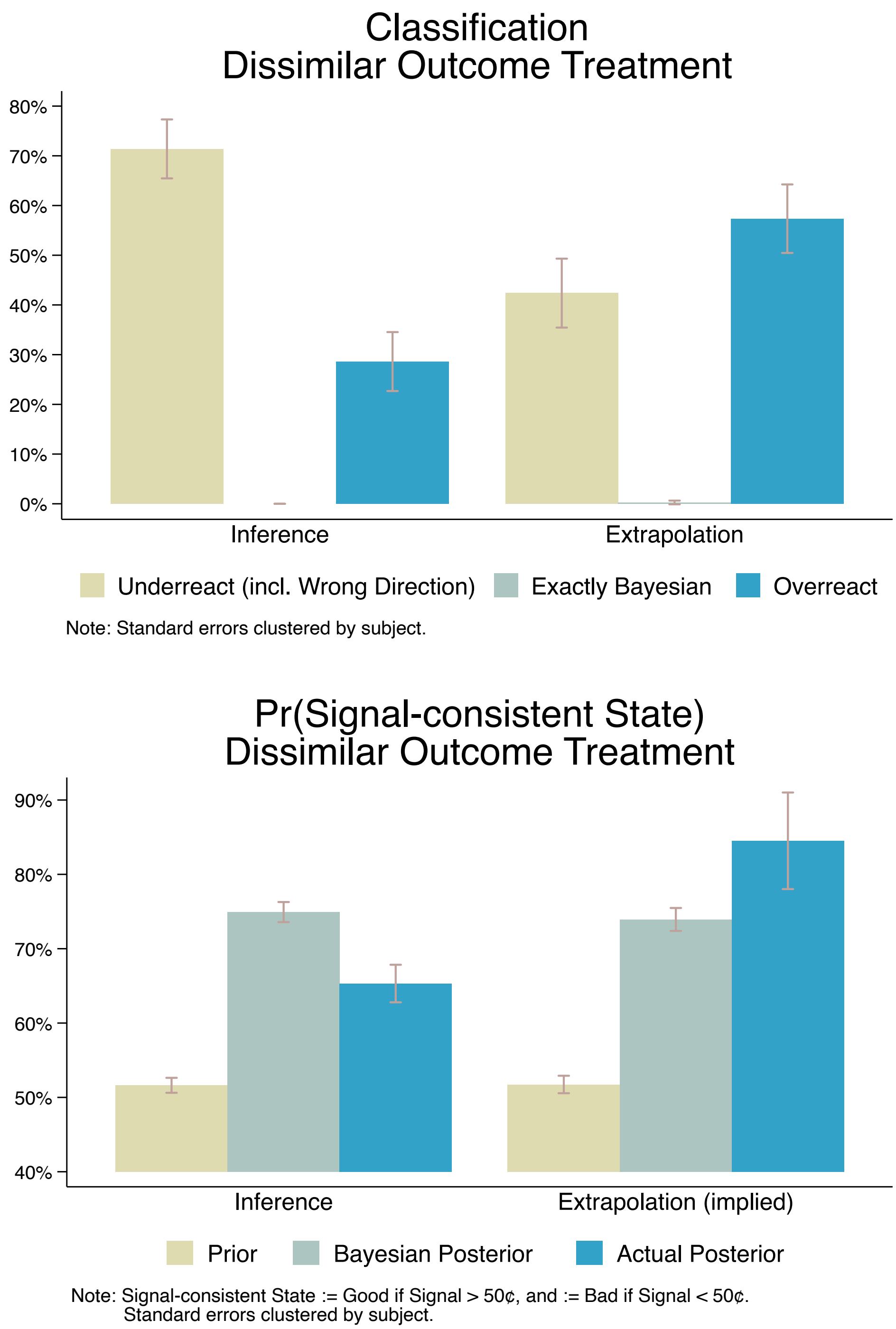


Figure 7: Classification and distribution of posteriors in the *Dissimilar Outcome* treatment.

Table 4: The gap between Inference and Extrapolation, classification results.

	Dependent Variable: Dummy for Over-reaction			
	All Problems		Problems with 50% Prior	
Extrapolation	0.256***	0.256***	0.265***	0.265***
	(0.039)	(0.040)	(0.041)	(0.044)
Problem FE	No	Yes	No	Yes
Subject FE	No	Yes	No	Yes
Observations	1388	1388	1066	1066
R^2	0.068	0.353	0.074	0.357

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 Treatment 1. Each observation corresponds either to an inference posterior in Part I1, or an extrapolation posterior in Part E1. We define the dependent variable as 1 if the posterior indicates that the subject updates from their subjective prior strictly more than what Bayesianism would imply, and as 0 otherwise. We drop problems in which the subject gives a prior outside $[0\%, 10\%]$ in Part E0, as the implied prior of the firm being good is then outside $[0\%, 100\%]$, and the implied Bayesian posterior cannot be properly defined. We also drop problems with completely uninformative signals ($m = 5\%$). In the third and fourth columns, we focus on problems where the true prior is 50%.

Table 5: The gap between Inference and Extrapolation in posterior beliefs.

	(Implied) Posterior: Pr(Signal-consistent State) (in $\%$)			
	All Problems		Problems with 50% Prior	
Extrapolation	17.780***	17.824***	17.451***	17.593***
	(3.652)	(3.808)	(4.174)	(4.391)
Implied Bayesian Posterior	0.469***	0.498***	0.438***	0.538***
	(0.065)	(0.074)	(0.113)	(0.133)
Problem FE	No	Yes	No	Yes
Subject FE	No	Yes	No	Yes
Observations	1388	1388	1066	1066
R^2	0.104	0.332	0.078	0.331

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 Treatment 1. Each observation corresponds either to an inference posterior in Part I1, or an extrapolation posterior in Part E1. We define the dependent variable as the (implied) posterior of the signal-consistent state, which is defined as Good if the signal $ishigherthan5\%, anddefinedasBadifthesignal$ is lower than 5%. *Implied Bayesian Posterior* is defined as the Bayesian posterior implied by the subject's own prior. We drop problems in which the subject gives a prior outside $[0\%, 10\%]$ in Part E0, as the implied prior of the firm being good is then outside $[0\%, 100\%]$, and the implied Bayesian posterior cannot be properly defined. We also drop problems with completely uninformative signals ($m = 5\%$). In the third and fourth columns, we focus on problems where the true prior is 50%.

Table 6: Problem-level classification of types.

Type	Extrapolation $E[s_1 \mid s_0]$	Inference $Pr(\text{Good} \mid s_0)$
Loading on the Signal-consistent State	20%	4%
Anchoring on the Signal	10%	3%
No Update	25%	29%
No Inference-Extrapolation Gap (excluding No Update)	6%	
Unclassified	44%	61%
N	1480	1480

Table 7: Subject-level classification of types.

Type	Extrapolation $E[s_1 \mid s_0]$	Inference $Pr(\text{Good} \mid s_0)$
Loading on the Signal-consistent State	13%	1%
Anchoring on the Signal	6%	1%
No Update	19%	24%
No Inference-Extrapolation Gap (excluding No Update)	0%	
Unclassified	60%	73%
N	202	202