Overreaction in Expectations: Evidence and Theory*

Hassan Afrouzi
Columbia

Spencer Y. Kwon
Harvard

Augustin Landier
HEC Paris

Yueran Ma
Chicago Booth

David Thesmar
MIT Sloan

September 20, 2021

Abstract

We investigate biases in expectations across different settings through a large-scale randomized experiment where participants forecast stable stochastic processes. The experiment allows us to control forecasters' information sets as well as the data generating process, so we can cleanly measure biases in beliefs. We report three facts. First, forecasts display significant overreaction to the most recent observation. Second, overreaction is stronger for less persistent processes. Third, overreaction is also stronger for longer forecast horizons. We develop a tractable model of expectations formation with costly processing of past information, which closely fits the empirical facts. We also perform additional experiments to directly test the mechanism of the model.

*This paper is a combination of two previous works "Biases in Expectations: Experimental Evidence" and "A Model of Costly Recall." Many thanks to Nick Barberis, Cary Frydman, Nicola Gennaioli, David Hirshleifer, Charlie Nathanson, Martin Schneider, Andrei Shleifer, Dmitry Taubinsky, Emanuel Vespa, and Mike Woodford, as well as conference participants at the AFA Annual Meeting and the Behavioral Approaches to Financial Decision Making Conference and seminar participants at Babson College, Berkeley-Chicago Behavioral Econ Workshop, Bocconi, BYU Marriott, Columbia, Duke, EDHEC, EIEF, the Fed, HEC Paris, Yale SOM, MIT Sloan, NBER, Northwestern Kellogg, NYU Stern, Singapore Management University, Stanford SITE, University of Amsterdam, USC Marshall, and Virtual Finance Workshop. David Norris provided very skillful research assistance on this project. Kwon is grateful for support from the Alfred P. Sloan Foundation Predoctoral Fellowship in Behavioral Macroeconomics, awarded through the NBER. Landier acknowledges financial support from the European Research Council under the European Community's Seventh Framework Program (FP7/2007-2013) Grant Agreement no. 312503 SolSys. This study is registered in the AEA RCT Registry and the identifying number is: "AEARCTR-0003173." Emails: hassan.afrouzi@columbia.edu, ykwon@hbs.edu, landier@hec.fr, yueran.ma@chicagobooth.edu, thesmar@mit.edu.

1 Introduction

Expectations formation plays a critical role in economics. A growing body of research using survey data shows that expectations often exhibit significant biases. Across different settings, however, the biases seem to vary. For instance, some studies document substantial overreaction, whereas others find less overreaction or some degree of underreaction.[1] These results raise an important question: why do biases in expectations vary across settings? Investigating this issue is a necessary step towards a systematic understanding of biases in expectations.

In this paper, we present new empirical evidence and theoretical analyses to illuminate how expectation biases vary with the persistence of the data generating process (DGP) and the forecast horizon. We begin with a large-scale randomized experiment to cleanly document the biases in expectations for different stochastic processes. Our experimental approach allows us to address three major concerns for studying expectations using survey data. First, we can control forecasters' information sets, which are not observable to the econometrician in survey data.[2] Second, we know the DGP and we can determine it, whereas the DGP is difficult for the econometrician to pin down in survey data. Finally, we can also control forecasters' payoff functions, whereas forecasters could have considerations other than accuracy in survey data. Overall, the experiment allows us to measure biases in forecasts in a precise way, trace out the structure of these biases and their variations across settings, and investigate whether commonly-used models align with the empirical evidence.

In our experiment, participants make forecasts of simple AR(1) processes. They are randomly assigned to a condition with a given AR(1) process with persistence ρ drawn from $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$; the mean is zero and the conditional volatility is 20. Participants observe 40 past realizations at the beginning and then make forecasts for another 40 rounds. In

[1]For overreaction in expectations, see De Bondt and Thaler (1990), Amromin and Sharpe (2013), Greenwood and Shleifer (2014), Gennaioli, Ma and Shleifer (2016), Bordalo, Gennaioli, La Porta and Shleifer (2019), Bordalo, Gennaioli, Ma and Shleifer (2020c), Barrero (2021), among others for evidence from forecasts of financial market and macroeconomic outcomes. For underreaction, see Abarbanell and Bernard (1992), Bouchaud, Krueger, Landier and Thesmar (2019), and Ma, Ropele, Sraer and Thesmar (2020) for evidence from forecasts of companies' near-term earnings.

[2]One workaround is to predict forecast errors using forecast revisions, since revisions are supposed to be within the forecaster's information set (Bordalo et al., 2020c). However, this approach has limitations, which we explain in detail in Section 2.2. Among other things, this method may be unreliable when the process is transitory, in which case the variance of forecast revisions may approach zero if beliefs are close to rational.

each round, participants observe a new realization and report one- and two-period-ahead forecasts. In follow-up experiments, we also extend the forecast horizon and elicit five- and ten-period-ahead forecasts.

We document three empirical facts. First, even though the process is simple and stable, rational expectations are strongly rejected in our data. In particular, forecasts in the data display significant overreaction to recent observations: the forecasts are systematically too high when the past realization is high, and vice versa. This feature is robust and it does not depend on whether participants know the process is AR(1), which we show using a sample of MIT students who understand AR(1) processes.

Second, we find that forecasts display more overreaction when the process is more transitory. This result echoes the patterns Bordalo et al. (2020c) observe in survey data. In the experiment, however, we can measure the degree of overreaction more precisely. Specifically, we calculate the persistence implied by participants' forecasts (i.e., the regression coefficient of the forecast $F_t x_{t+1}$ on x_t) and compare it with the actual persistence of the process. In our setting, the gap between the implied persistence and the actual persistence provides a clear measure of overreaction. In the data, the implied persistence is close to one when the process is a random walk. When the actual persistence decreases, the implied persistence decreases but less than one for one, so the gap between implied and actual persistence is larger when the process is more transitory. For instance, the implied persistence is 0.85 when the actual persistence is 0.6 and 0.45 when the actual process is i.i.d. (i.e., more overreaction when the process persistence is lower).

Third, we find that overreaction is stronger for longer-horizon forecasts. The forecast-implied persistence (per period) is higher for long-horizon forecasts than for short-horizon forecasts, based on forecasts the same participants made for the same AR(1) process.[3] Accordingly, biases arising from forecasters using a given incorrect value of the persistence parameter (e.g., Gabaix, 2018) is not sufficient to account for the behavior of forecasts across different horizons. Our finding aligns with the evidence in Giglio and Kelly (2018), who show that affine asset pricing models using a given process persistence cannot simultaneously account for the price movements of short-maturity and long-maturity claims, with

[3]For general forecast horizons, we compute the implied persistence per period by taking the regression coefficient of the forecast $F_t x_{t+h}$ on x_t and raising it to the $1/h$-th power.

"much higher persistence implied from the long end."

We then examine whether commonly-used expectations models account for the empirical facts. We find that standard formulations of these models do not perform well. For instance, the older adaptive or extrapolative models generate forecast-implied persistence that does not vary much with the actual persistence, so overreaction in these models is too strong for more transitory processes. In contrast, more recent models such as constant gain learning of the process (Evans and Honkapohja, 2001) and diagnostic expectations (Bordalo, Gennaioli and Shleifer, 2018) generate implied persistence that varies too much with the actual persistence, so overreaction is too weak for more transitory processes. The standard specification of these models generates minimal or no overreaction when the process is i.i.d., as they do not allow transient shocks to have much impact on expectations.

Building on Nagel and Xu (2019) and da Silveira and Woodford (2019), we find that a parsimonious way to account for the evidence is to allow recent observations to influence the assessment of the long-run mean of the process. We provide a tractable model that microfounds the degree of overreaction, yields closed-form solutions that are easy to estimate in the data, and generates additional testable predictions. In the model, the agent processes information to form beliefs regarding the long-run mean at each point in time, where recent information is less costly to process. This mechanism connects with the literature in psychology which emphasizes that some information, especially more recent information, is in a state of "heightened activation," and is thus more actively utilized than others (Baddeley and Hitch, 1993; Cowan, 1998, 2017a). Such heightened activation applies to a wide range of information processing settings and is not limited to explicit recall.[4] We refer to the set of information actively processed as what is "on top of mind."[5] The model nests the rational benchmark in the frictionless limit where all information is on top of mind. Otherwise, the model delivers a partial dependence of the long-run mean assessment on the recent observation which generates greater overreaction when the process is less persistent. The model also predicts greater overreaction for longer forecast horizons, and closely fits the term structure

[4]For instance, the survey by Barrett, Tugade and Engle (2004) notes that "Any environment contains an array of stimuli simultaneously ... stronger representations will laterally inhibit weaker ones, and the strongest will be expressed in behavior." We discuss this in more detail in Section 5.1.

[5]The set of information that is actively processed has been referred to in the psychology literature as heightened activation, increased accessibility, conscious awareness, or focus of attention, and is connected more broadly to costly information processing in a variety of settings (Spillers, Brewer and Unsworth, 2012).

of forecast biases as non-targeted moments.

We take our model to the data by minimizing the mean squared error with respect to all one-period-ahead forecasts in our baseline experiment (we test other models in the literature in the same way). We calculate the implied persistence in the experimental data and the value generated by the model. The implied persistence based on our model closely matches that in the data for all values of ρ. We then use the model estimated on one-period-ahead forecasts to compute the implied persistence for long-horizon forecasts as non-targeted moments. For forecast horizons of two, five, and ten covered by our experiments, our model closely matches the degree of overreaction observed in the data for all values of ρ.

The key mechanism in our model is that individuals react more to information that is more on top of mind. To test this mechanism directly, we design additional experiments that aim to influence what is on top of mind by drawing participants' focus away from the most recent observation. In one condition, we require participants to click on the realization ten periods ago before they can make new forecasts in each round. In another condition, we show a red line at zero in the experimental interface. In both cases, the degree of overreaction decreases relative to that in the baseline condition, in line with predictions of our model.

Finally, we provide corroborating evidence using data from financial markets, which shows the external validity and relevance of our findings. Using data on stock analysts' forecasts, we find that overreaction in forecasts of firms' sales growth is stronger when sales growth is less persistent. Using data on asset prices, we show that the "value premium" (i.e., companies with a high book-to-market ratio tend to have higher stock returns), which is often viewed as a reflection of overreaction, is also stronger among firms with less persistent sales growth. In sum, we believe that biases in expectations are not random; the findings we document add to a systematic understanding of how the degree of overreaction varies across settings, and how results from various studies can be connected.

Literature Review. Our work is related to three branches of literature. First, our empirical findings complement recent studies using survey data discussed in the first paragraph, which document strong overreaction in some settings and weaker overreaction or underreaction in others. As mentioned before, while analyses using survey data are highly valuable, they face major obstacles given that researchers do not know forecasters' information sets, payoff functions, and the DGP. A key contribution of our paper is implementing a large-scale experiment to cleanly connect biases in expectations with both the properties of the underlying process and the forecast horizon.

Second, we contribute to experimental studies of forecasts (see Assenza, Bao, Hommes and Massaro (2014) for a survey). Prior work on forecasting stochastic processes includes Hey (1994), Frydman and Nave (2016) and Beshears, Choi, Fuster, Laibson and Madrian (2013); we provide an extensive review of this literature in Table A.1. Most closely related, Reimers and Harvey (2011) also document that the forecast-implied persistence is higher than the actual persistence for transitory processes, which indicates the robustness of this phenomenon, but they do not analyze the term structure of forecasts or test models of expectations. Overall, relative to existing research, our experiments have a large scale, a wide range of settings, and diverse demographics; we also collect the term structure of forecasts. In addition, we use the data to investigate a number of commonly-used models, while prior studies tend to focus on testing a particular type of model.[6]

Finally, we use our data to shed further light on models of expectation formation, highlighting how biases in the assessment of the long-run mean can be a parsimonious way to account for key empirical facts. Some modeling techniques we use are related to the literature on noisy perception and rational inattention (Woodford, 2003; Sims, 2003). This literature has focused on frictions in perception of new information. Instead, our model emphasizes frictions in exploiting past information, which is key for generating overreaction. Another set of models postulate that forecasters use an incorrect value of the persistence ρ (Gabaix, 2018; Angeletos, Huo and Sastry, 2021). We find that a given "mistaken" ρ cannot simultaneously account for the degree of overreaction in short-term and long-term forecasts. If using an incorrect ρ is the main bias, overreaction will dissipate for long-term forecasts, which is not the case in the data.

Several recent models examine the role of memory in belief formation, which also feature frictions in exploiting past information. Bordalo, Gennaioli and Shleifer (2020b), Bordalo, Coffman, Gennaioli, Schwerter and Shleifer (2020a), and Bordalo, Conlon, Gennaioli, Kwon

[6] Some experimental studies document underreaction in inference problems about an underlying state (e.g., assess the urn type after a red or blue ball is drawn from the urn), which Benjamin (2019) summarized as "underinference." Recent experimental findings by Fan, Liang and Peng (2021) show that people simultaneously underreact in the inference problem (e.g., inferring whether a company is good or bad after observing a positive stock return) and overreact in the forecasting problem (e.g., predicting the stock return in the next period). This evidence suggests that people approach the inference problem differently from the forecasting problem.

and Shleifer (2021) build on representativeness (Kahneman and Tversky, 1972) and associative recall (Kahana, 2012). Wachter and Kahana (2021) present a retrieved-context theory of beliefs to model associative recall.[7] The most closely related analysis here is da Silveira, Sung and Woodford (2020), which generalizes earlier work by da Silveira and Woodford (2019). They present a dynamic model of optimally noisy memory where past information is summarized by a memory state formed before each period; when the memory is imprecise, the agent optimally puts more weight on the latest observation, which generates overreaction. In our model, the over-weighting of recent information can be related to memory constraints, but it can also arise from other frictions in information processing that lead to heightened activation and disproportionate focus on recent information (see, e.g., Baddeley and Hitch, 1993; Spillers, Brewer and Unsworth, 2012; Cowan, 2017a), which we discuss in more detail in Section 5.1.

Although we focus on overreaction given our empirical findings, we provide an extension of our model in Appendix D which accommodates underreaction by introducing noisy signals to the belief formation process. These noisy signals can play a role in survey data (Coibion and Gorodnichenko, 2012, 2015; Kohlhas and Walther, 2021), but are unlikely to be first-order in our simple forecasting experiment (so overreaction dominates here). In this extension, the relative degree of overreaction is still stronger when the process is less persistent, consistent with our evidence and findings from field data in Sections 2.1 and 6.2. Meanwhile, the average level of the bias can be overreaction or underreaction, depending on the noisiness of signals.

The rest of the paper proceeds as follows. Section 2 discusses stylized facts from survey data and the limitations of these analyses, which motivate our experiment. Section 3 describes the experiment. Section 4 presents our main empirical facts that overreaction is stronger for less persistent processes and longer forecast horizons. It also analyzes whether commonly-used models align with the evidence. Section 5 presents our model and shows that the model fits the data well. Section 6 performs robustness checks of our model and documents additional evidence from asset prices in financial markets. Section 7 concludes.

[7] In addition, Nagel and Xu (2019) and Neligh (2020) study applications of memory decay. In empirical analyses, Enke, Schwerter and Zimmermann (2021) experimentally test how associative recall affects beliefs. Hartzmark, Hirshman and Imas (2021) and D'Acunto and Weber (2020) also find evidence consistent with memory playing a role in decision making.

2 Motivating Facts

To motivate our experimental investigation, we first present stylized facts from survey forecasts and discuss their limitations.

2.1 Variations in Overreaction: Evidence from Survey Forecasts

A major challenge for analyzing expectations using survey forecast data is that the true DGP and forecasters' information sets are both unknown. Taking inspiration from Coibion and Gorodnichenko (2015), Bordalo et al. (2020c) observe that one idea is to capture belief updating using forecast revisions by individual forecasters: revisions should incorporate news that a forecaster responds to and should be part of the information set. When a forecaster overreacts to information, revisions at the individual level would over-shoot (e.g., upward forecast revisions would predict realizations below forecasts). The empirical specification is the following, which regresses forecast errors on forecast revisions in a panel of quarterly individual-level forecasts:

$$\underbrace{x_{t+h} - F_{i,t}x_{t+h}}_{\text{forecast error}} = a + b\underbrace{\left(F_{i,t}x_{t+h} - F_{i,t-1}x_{t+h}\right)}_{\text{forecast revision}} + v_{it}, \tag{2.1}$$

where $F_{i,t}x_{t+1}$ is the forecast of individual i of outcome x_{t+h}. For each series, we obtain a coefficient b (henceforth the "error-revision coefficient"). When overreaction is present, b should be negative, and vice versa (Bordalo et al., 2020c).

Bordalo et al. (2020c) analyze professional forecasts of 22 series of macroeconomic and financial variables. They find that the error-revision coefficient b is generally negative, and it is more negative for processes with lower persistence. They interpret this pattern as an indication that overreaction tends to be stronger when the actual process is more transitory. In Figure I, Panel A, we use Survey of Professional Forecasters (SPF) data and replicate this finding. Here we use the simple one-period-ahead forecasts, namely $h = 1$. The y-axis shows the coefficient b for different series, and the x-axis shows the autocorrelation of each series as a simple measure of persistence. We see that the coefficient b is more negative (i.e., overreaction is stronger) when the actual series is less persistent.

In Figure I, Panel B, we also document similar results using analysts' forecasts of firms' sales from the Institutional Brokers' Estimate System (IBES). Again we use one-period-ahead forecast, namely $h = 1$. We normalize both actual sales and projected sales using lagged sales, and the frequency is quarterly. Results are very similar if we use an annual frequency, or using earnings forecasts instead of sales forecasts.[8] We run one regression in the form of Equation (2.1) for each firm i to obtain coefficient b_i. We also compute the autocorrelation of the actual sales growth process ρ_i. Figure I, Panel B, shows a binscatter plot of the average b_i in twenty bins of ρ_i. Here the key fact remains: the coefficient b_i is more negative when the actual sales process of the firm is less persistent. Nonetheless, the average level of b_i is positive (indicating underreaction), consistent with Bouchaud et al. (2019). In contrast, Bordalo et al. (2019) find negative error-revision coefficients (pointing to overreaction) when analyzing analyst forecasts about long-term growth instead of near-term cash flows, which suggests the forecast horizon may affect the degree of overreaction as well.

2.2 Challenges in Field Data

Although these results from survey data are intriguing, they can be difficult to interpret unequivocally for several reasons.

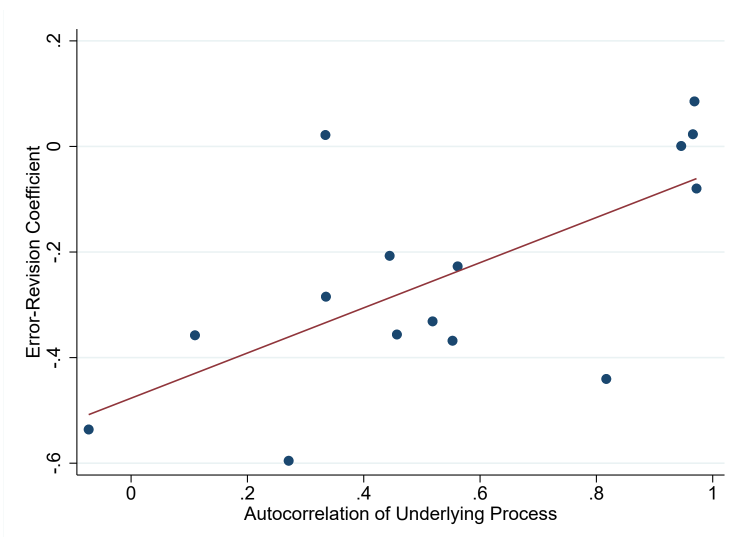
First, the error-revision regressions have limitations. To begin, it is difficult to estimate b precisely for transitory processes when expectations are close to rational. In this case, revisions are close to zero, so the regression coefficient is not well estimated. As an illustration, in Figure A.1, Panel A, we show the error-revision coefficient b from simulations where we simulate forecasters under diagnostic expectations (Bordalo et al., 2018, 2020c) for AR(1) processes with different levels of persistence. By construction, the simulated coefficient (shown by the solid line) is on average similar to theoretical predictions in the diagnostic expectations model (Bordalo et al., 2020c). Meanwhile, the dashed lines show that the confidence

[8] Earnings forecasts have several complications relative to sales forecasts. First, earnings forecasts primarily take the form of earnings-per-share (EPS), which may change if firms issue/repurchase shares, or have stock splits/reverse splits. This requires us to transform EPS forecasts to implied forecasts about total firm earnings, which could introduce additional measurement error. Second, the definition of earnings firms use for EPS can be informal ("pro forma" earnings, instead of formal net income according to the Generally Accepted Accounting Principles (GAAP). As a result, matching earnings forecasts properly with actual earnings can be more challenging. In comparison, sales forecasts are directly about total sales of the firm, and the accounting definition of sales is clear (based on GAAP).

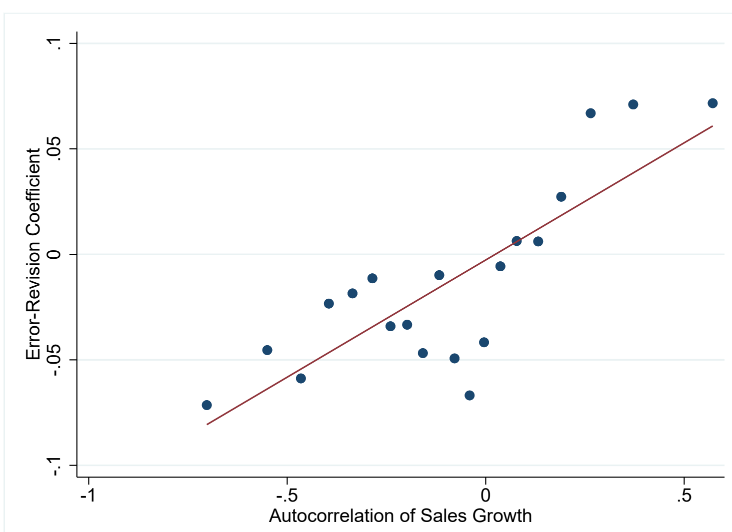
Figure I: Forecast Error on Forecast Revision Regression Coefficients

In Panel A, we use SPF data on macroeconomic forecasts and estimate a quarterly panel regression using each individual j's forecasts for each variable x_i: $x_{i,t+1} - F_{i,j,t}x_{i,t+1} = a + b_i(F_{i,j,t}x_{i,t+1} - F_{i,j,t-1}x_{i,t+1}) + v_{i,j,t}$, where the left-hand-side variable is the forecast error and the right-hand-side variable is the forecast revision for each forecaster j. The y-axis plots the regression coefficient b_i for each variable, and the x-axis plots the autocorrelation of the variable. The variables include quarterly real GDP growth, nominal GDP growth, GDP price deflator inflation, CPI inflation, unemployment rate, industrial production index growth, real consumption growth, real nonresidential investment growth, real residential investment growth, real federal government spending growth, real state and local government spending growth, housing start growth, unemployment rate, 3-month Treasury yield, 10-year Treasury yield, and AAA corporate bond yield. In Panel B, we use IBES data on analyst forecasts of firms' sales and estimate a quarterly panel regression using individual analyst j's forecasts for each firm i's sales $x_{i,t+1} - F_{i,j,t}x_{i,t+1} = a + b_i(F_{i,j,t}x_{i,t+1} - F_{i,j,t-1}x_{i,t+1}) + v_{i,j,t}$, where the left-hand-side variable is the forecast error and the right-hand-side variable is the forecast revision for each forecaster j. The y-axis plots the regression coefficient b_i, and the x-axis plots the autocorrelation of firm i's sales. For visualization, we group firms into twenty bins based on the persistence of their sales, and present a binscatter plot. Both actual and projected sales are normalized by lagged book assets.

Panel A. SPF Forecasts



Panel B. Analyst Forecasts



intervals become very wide when the process persistence is below 0.5.[9] The intuition in this example is that the variance of the right-hand-side variable, the forecast revision, goes to zero for i.i.d. processes when expectations are close to rational (see discussion on asymptotic standard errors in Appendix B.1).

Estimation issues aside, the error-revision coefficient b is not necessarily a direct metric for the degree of overreaction (i.e., how much subjective beliefs over-adjust relative to the rational benchmark). This empirical coefficient does not directly map into a structural parameter, and its interpretation can be model dependent.[10] In addition, the error-revision coefficient b can be subject to the critique that if the forecast $F_t x_{t+h}$ is measured with noise, the regression coefficient b could be mechanically negative, given that $F_t x_{t+h}$ affects both the right side (forecast revision) and the left side (forecast error) of the regression. Ultimately, as explained earlier, one important reason for using the error-revision regression in survey forecast data is that researchers do not observe forecasters' information sets, and forecast revision can be used to capture information they respond to. If the information set is clear, however, then we can directly analyze forecasts and errors using components of the information sets.

Second, the DGP can be difficult to pin down in survey data. Many models assume the DGP to be AR(1), and the interpretation of the regression coefficients (such as Equation (2.1)) can change if the DGP is not AR(1).[11] However, econometricians may not be able to statistically differentiate whether a process is AR(1) or ARMA with longer lags in finite sample (Fuster, Laibson and Mendel, 2010).

These challenges in the field data show that complementary experimental analyses would be useful. Accordingly, we implement a large-scale forecast experiment where the forecasting environment is simple and the DGP is clearly defined, which allows us to measure over/underreaction precisely. The experiments also allow us to randomly assign participants into different conditions, so we can cleanly document the properties of subjective

[9]For AR(1) processes, the diagnostic forecast is $E_t^\theta x_{t+1} = E_t x_{t+1} + \rho \epsilon_t$, where $E_t x_{t+1}$ is the rational forecast, ρ is the AR(1) persistence, and ϵ_t is the news in period t. When the process is i.i.d., the diagnostic forecast becomes the same as the rational forecast, and the error-revision coefficient is not well-defined.

[10]In particular, since the forecast revision in period t is the change between the subjective forecast from $t-1$ to t ($F_t x_{t+h} - F_{t-1} x_{t+h}$), its size and variance are affected by the past forecast ($F_{t-1} x_{t+h}$), so the magnitude of the error-revision coefficient b can be path dependent.

[11]For instance, Bordalo et al. (2020c) show that the regression specification in Equation (2.1) no longer holds and a modified specification is required if the DGP is AR(2) not AR(1).

forecasts in different settings.

In the experiment, for AR(1) processes a straightforward way to measure the degree of over/underreaction is to examine the sensitivity of subjective forecasts to realized observations. We can regress the forecast $F_t x_{t+h}$ on x_t to obtain coefficient $\rho_{s,h}$, and then calculate $\rho^s_h = \rho^{1/h}_{s,h}$ as the implied persistence of the forecasts. We can compare the implied persistence ρ^s_h with the actual persistence ρ of the process. When $\rho^s_h > \rho$, there is overreaction, in the sense that the forecast displays excess sensitivity to the latest observation x_t (i.e., when x_t is high, the forecast tends to be too high, and vice versa). Relative to the error-revision coefficient, the magnitude of the implied persistence (ρ^s_h) is easier to interpret as a measure of overreaction: the gap between ρ^s_h and ρ directly captures how much more forecasts respond to recent observations relative to the rational benchmark.[12] In addition, this approach does not face the econometric complications that apply to the error-revision coefficient we discussed above, as the variance of the right-hand-side variable (namely the past realization) is always well-defined (it does not vanish to zero as ρ decreases). Figure A.1, Panel B, presents simulations of the implied persistence (with $h = 1$) and shows this approach is reliable for all levels of persistence.

We explain the experimental design in the next section. Our experiments focus on AR(1) processes for three reasons based on the previous discussion. First, the rational benchmark is clear in this setting. Second, we can precisely measure the degree of over-/under-reaction relative to the benchmark. Third, most models of expectations formation also focus on AR(1), so we can compare our findings with their predictions.

3 Experiment Design

We design a simple forecasting experiment, where the DGP is an AR(1) process:

$$x_{t+1} = (1-\rho)\mu + \rho x_t + \epsilon_t. \tag{3.1}$$

[12] There is an approximate relationship between the error-revision coefficient and $\zeta(\rho,h) = (\rho^s_h)^h/\rho^h$. Specifically, $1/(1+b) = \frac{Var(FR)}{Cov(FE+FR,FR)}$. If we set $F_{t-1}x_{t+h}$ as a constant, then this coefficient is the same as $\zeta(\rho,h)$. Accordingly, a negative error-revision coefficient, often interpreted as evidence of overreaction, implies $\zeta(\rho,h) > 1$, i.e., overreaction of the subjective belief to the latest observation.

The experiment begins with a consent form, followed by instructions and tests. Participants first observe 40 past realizations of the process. Then, in each round, participants make forecasts and observe the next realization, for 40 rounds. After the prediction task, participants answer some basic demographic questions.

Each participant is only allowed to participate once. Participants include both individuals across the US from Amazon's online Mechanical Turk platform (MTurk) and MIT undergraduates in Electrical Engineering and Computer Science (EECS). For MTurk, we use HITs titled "Making Statistical Forecasts."[13] For MIT students, we send recruiting emails to all students with a link to the experimental interface.

3.1 Experimental Conditions

There are four main sets of experiments, which we describe below and summarize in Table A.2 in the Appendix.

Experiment 1 (Baseline, MTurk). Experiment 1 is our baseline test, conducted in February 2017 on MTurk. We use six values of ρ: $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$. The volatility of ϵ is 20. The constant μ is zero. Participants are randomly assigned to one value of ρ. Each participant sees a different realization of the process. At the beginning, participants are told that the process is a "stable random process." In each round, after observing realization x_t, participants predict the value of the next two realizations x_{t+1} and x_{t+2}. Figure A.2 provides a screenshot of the prediction page. There are 207 participants in total and about 30 participants per value of ρ.

Experiment 2 (Long Horizon, MTurk). Experiment 2 investigates longer horizon forecasts. We assign participants to conditions identical to Experiment 1, except that we collect forecasts of x_{t+1} and x_{t+5} (instead of x_{t+2}), with $\rho \in \{0.2, 0.4, 0.6, 0.8\}$. Experiment 2 was conducted in June 2017 on MTurk. There are 128 participants in total.

Experiment 3 (Describe DGP, MIT EECS). In Experiment 3, we study whether provid-

[13]The MTurk platform is commonly used in experimental studies (Kuziemko, Norton, Saez and Stantcheva, 2015; D'Acunto, 2015; Cavallo, Cruces and Perez-Truglia, 2017; DellaVigna and Pope, 2017, 2018). It offers a large subject pool and a more diverse sample compared to lab experiments. Prior research also finds the response quality on MTurk to be similar to other samples and to lab experiments (Casler, Bickel and Hackett, 2013; Lian, Ma and Wang, 2018).

ing more information about the DGP affects forecasts. To make sure that participants have a good understanding of the AR(1) formulation, we perform this test among MIT undergraduates in Electrical Engineering and Computer Science (EECS). Experiment 3 was conducted in March 2018 and there are 204 participants. We use the same structure as in Experiment 1, with AR(1) persistence $\rho \in \{0.2, 0.6\}$. For each persistence level, the control group is the same as Experiment 1, and the process is described as "a stable random process." For the treatment group, we describe the process as "a fixed and stationary AR(1) process: $x_t = \mu + \rho x_{t-1} + e_t$, with a given μ, a given ρ in the range [0,1], and e_t is an i.i.d. random shock." Thus there are $2 \times 2 = 4$ conditions in total, and participants are randomly allocated to one of them. At the end of the experiment, we further ask students questions testing their prior knowledge of AR(1) processes.[14]

Experiment 4 (Additional Test, MTurk). In Experiment 4, we study how changing the focus of participants affects the results. We have a baseline condition that is the same as Experiment 1 and two treatment conditions with different designs discussed in more detail in Section 5.5. In addition, we have a condition where participants forecast x_{t+1} and x_{t+10}. All conditions have $\rho \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. Participants are randomly assigned into a given condition and a given level of ρ. As before, there are about 30 participants for each treatment condition and level of ρ. Experiment 4 was conducted in March 2021 on MTurk.

3.2 Payments

We provide fixed participation payments as well as incentive payments that depend on the performance in the prediction task. For the incentive payments, participants receive a score for each prediction that increases with the accuracy of the forecast (Dwyer, Williams, Battalio and Mason, 1993; Hey, 1994): $S = 100 \times \max(0, 1 - |\Delta|/\sigma)$, where Δ is the difference between the prediction and the realization, and σ is the volatility of the noise term ϵ. This loss function ensures that a rational participant will optimally choose the rational expectation, and it ensures that payments are always non-negative. A rational agent would expect

[14]We do not disclose the values of μ and ρ, since the objective of our study is to understand how people form forecasting rules; directly providing the values of μ and ρ would make this test redundant.

to earn a total score of about 2,800.[15] We calculate the cumulative score of each participant, and convert it to dollars. The total score is displayed on the top left corner of the prediction screen (see Figure A.2).

For experiments on MTurk (Experiments 1, 2, and 4), the base payment is \$1.8; the conversion ratio from the score to dollars is 600, which translates to incentive payments of about \$5 for rational agents. For experiments with MIT students (Experiment 3), the base payment is \$5; the conversion ratio from the score to dollars is 240, which translates to incentive payments of about \$12 for rational agents.

3.3 Summary Statistics

Table A.3 shows participant demographics and other experimental statistics. Overall, MTurk participants are younger and more educated than the U.S. population. The mean duration of the experiment is about 18 minutes, and the hourly compensation is in the upper range of tasks on MTurk. As expected, MIT EECS undergrads are younger. Their forecast duration and overall forecast scores are similar to the MTurk participants.[16]

4 Main Empirical Findings

In this section, we present the main findings from our experiments. In Section 4.1, we present the key empirical facts. In Section 4.2, we then analyze whether commonly-used models of expectations are in line with these facts.

As mentioned in Section 2, the error-revision regression approach has limitations for accurately measuring the degree of overreaction. In our experiment, a natural and more

[15] $E(1 - |x_{t+1} - F_t|/\sigma)$ is maximal for a forecast F_t equal to the 50th percentile of the distribution of x_{t+1} conditional on x_t. Given that our process is symmetric around the rational forecast, the median is equal to the mean, and the optimal forecast is equal to the conditional expectation. Whether the rational agent knows the true ρ (Full Information Rational Expectations) or predicts realizations using linear regressions (Least-Square Learning) does not change the expected score by much. In simulations, over 1,000 realizations of the process, we find that expected scores of the two approaches differ by less than .3%.

[16] The participation constraint is likely to be satisfied. For the MTurk tests, the average realized total payment (participation plus incentive payment) is about \$5 (for a roughly 15 minute task), which is high compared to the average pay rate. For the MIT tests, the average realized total payment is around \$15. The payments are sufficiently attractive to recruit 200 EECS undergrads out of 1,291 students within six hours. For the incentive compatibility constraint, recent work by DellaVigna and Pope (2017) show that participants provide high effort even when the size of the incentive payment is modest, and the power of incentives does not appear to be a primary issue in this setting.

precise alternative measure of the degree of overreaction is the persistence implied by the forecast. Specifically, denote $\rho_{s,h}$ as the coefficient in the regression:

$$F_{it}x_{t+h} = c + \rho_{s,h}x_t + u_{it}, \tag{4.1}$$

for each level of AR(1) persistence ρ and forecast horizon h (estimated in the panel of individual-level forecasts).[17] The implied persistence is then given by $\rho_h^s = \rho_{s,h}^{1/h}$. As the Full Information Rational Expectation (FIRE) is given by $\rho^h x_t$, the difference between ρ_h^s and ρ provides a direct measure of the degree of overreaction. This measure is reliable for AR(1) processes as we show in Section 2, and forecasters' information sets are clear in the experiment.

4.1 Basic Facts: More Overreaction for More Transitory Processes and Longer Forecast Horizons

A. Overreaction and Process Persistence

We first describe how overreaction varies for AR(1) processes with different levels of persistence, starting with one-period-ahead forecasts. Using data from Experiment 1, we have AR(1) processes with persistence from 0 to 1. First, in Figure II, Panel A, we run the error-revision regression in Equation (2.1), as we did using field data (the y-axis shows the error-revision coefficient, and the x-axis shows the persistence of the process). We find that the coefficient b is more negative for transitory processes, in line with the evidence from field data discussed in Section 2. Second, in Figure II, Panel B, we plot the implied persistence ρ_1^s against the actual process persistence ρ. We see that when $\rho = 1$, ρ_1^s is around one (i.e., the subjective and rational forecasts have roughly the same sensitivity to x_t). When ρ is smaller, ρ_1^s declines, but not as much. When $\rho = 0$, ρ_1^s is around 0.45 (i.e., the sensitivity of the subjective forecast to x_t is much larger than that under the rational benchmark).[18]

Overall, in the experiment, by explicitly controlling the DGP and forecasters' information

[17] We can also run the regression for every individual forecaster and take the mean or median regression coefficient for each process persistence ρ and forecast horizon h. The results are similar.

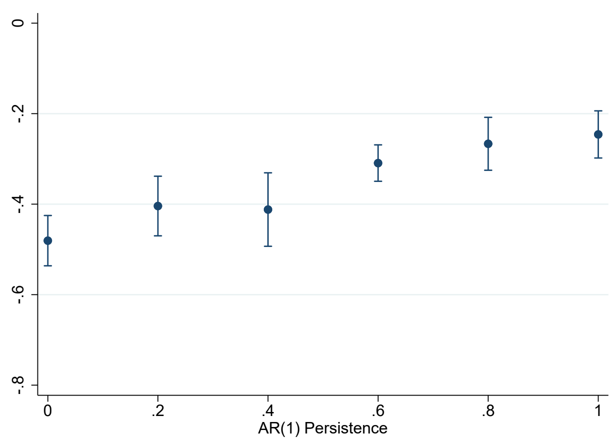
[18] We can also compute the ratio of relative overreaction $\zeta(\rho, h) = \frac{\rho_{s,h}}{\rho^h}$. Figure A.3 plots the value of $\zeta(\rho, h)$ for each level of ρ (except when $\rho = 0$ where $\zeta(\rho, h)$ is not well-defined). Since ρ_1^s decreases less than one-for-one with ρ, the degree of overreaction is higher when the process is less persistent.

sets, we can establish clearly that overreaction is stronger for more transitory processes.

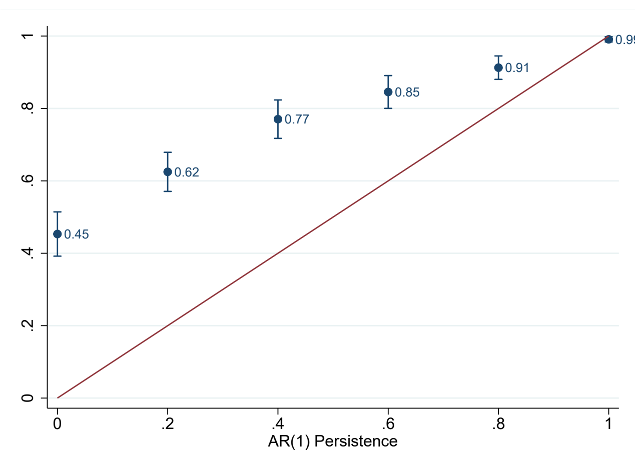
Figure II: Implied Persistence and Actual Process Persistence

In Panel A, we use data from Experiment 1 and for each level of AR(1) persistence ρ, we estimate a panel regression of forecast errors on forecast revisions: $x_{t+1} - F_{i,t}x_{t+1} = a + b(F_{i,t}x_{t+1} - F_{i,t-1}x_{t+1}) + v_{it}$. The y-axis plots the regression coefficient b, and the x-axis plots the AR(1) persistence ρ. In Panel B, we estimate the implied persistence ρ_1^s from $F_{it}x_{t+1} = c + \rho_1^s x_t + u_{it}$ for each level of AR(1) persistence ρ. The y-axis plots the implied persistence ρ_1^s, and the x-axis plots the AR(1) persistence ρ. The red line is the 45-degrees line, and corresponds to the implied persistence under Full Information Rational Expectations (FIRE). The vertical bars show the 95% confidence interval of the point estimates.

Panel A. Forecast Error on Forecast Revision Regression Coefficients



Panel B. Forecast-Implied Persistence and Actual Persistence



FIRE vs. In-Sample Least Square Learning. The analyses above use FIRE for the actual persistence ρ. Results are similar if we use in-sample least square learning as the rational benchmark instead. Specifically, the in-sample least square estimates are formed as:

$$\widehat{E}_t x_{t+h} = \widehat{a_{t,h}} + \sum_{k=0}^{k=n} \widehat{b_{k,h,t}} x_{t-k}. \tag{4.2}$$

In period t the forecaster predicts x_{t+h} using lagged values from x_{t-k} up to x_t; parameters $\widehat{a_{t,h}}$ and $\widehat{b_{k,h,t}}$ are estimated, on a rolling basis, using OLS and past realizations until x_t. The estimated coefficients may differ based on persistence ρ. We set $n = 3$, but results are not sensitive to the number of lags.

In our data, the difference between $\widehat{E}_t x_{t+h}$ and FIRE is small. The top panel of Figure A.4 shows that the mean squared difference between these two expectations is small, and does not decrease much after 40 periods. This is because our AR(1) processes are very simple, and a few dozen data points are enough for least square forecasts to be reasonably accurate. It also shows that the mean squared difference between the least square forecast and the actual forecasts are substantial, and does not change much across different periods. The bottom panel shows that the persistence implied by least square learning is about the same as the actual persistence ρ. Accordingly, in the rest of the paper we use FIRE in our baseline definitions, but all the results are similar if we use the in-sample least square $\widehat{E}_t x_{t+h}$ instead.

Effect of Linear AR(1) Prior. We also analyze whether explicitly providing a prior that the DGP is a linear AR(1) process affects the results. In Experiment 1 with participants from the general population, we describe the process as a "stable random process" (given that most of these participants may not know what an AR(1) process means). In Experiment 3 with MIT EECS students, we tell half of the participants that the DGP is AR(1) with fixed μ and ρ (treatment group), and half of the participants the process is a "stable random process" (control group). Given the size limit of MIT EECS students, we use two values of ρ: 0.2 and 0.6 (so participants are randomly assigned into a given value of rho and a given type of process description).

In Figure A.5, we show that whether the linear AR(1) information was provided has no discernible impact on the properties of forecast errors. In Panel A, we plot the distributions of the forecast errors, which are almost identical in the treatment vs. control group. In Panel B, we find that the predictability of forecast errors conditional on the latest observation x_t is also similar in the treatment vs. control group. In both samples, forecasts tend to be too high when x_t is high (overreaction), and the magnitude of the bias is about the same. Table A.4 shows that the implied persistence is also similar in both the treatment and control groups. Overall, we find that explicit descriptions of the AR(1) process do not seem to affect the basic patterns in the data. Put differently, participants do not seem to enter the experiment with complicated nonlinear priors. Finally, Figure A.6 compares the implied persistence ρ_1^s in the MIT experiments with that in our baseline experiments, which shows the results are very stable (even with the caveat that these two sets of experiments were not conducted and randomized at the same time).

Stability across Demographics. Figure A.7 shows both the error-revision coefficient b and implied persistence ρ_1^s against ρ in different demographic groups. In all cases, the main patterns are stable.

B. Overreaction and Forecast Horizon

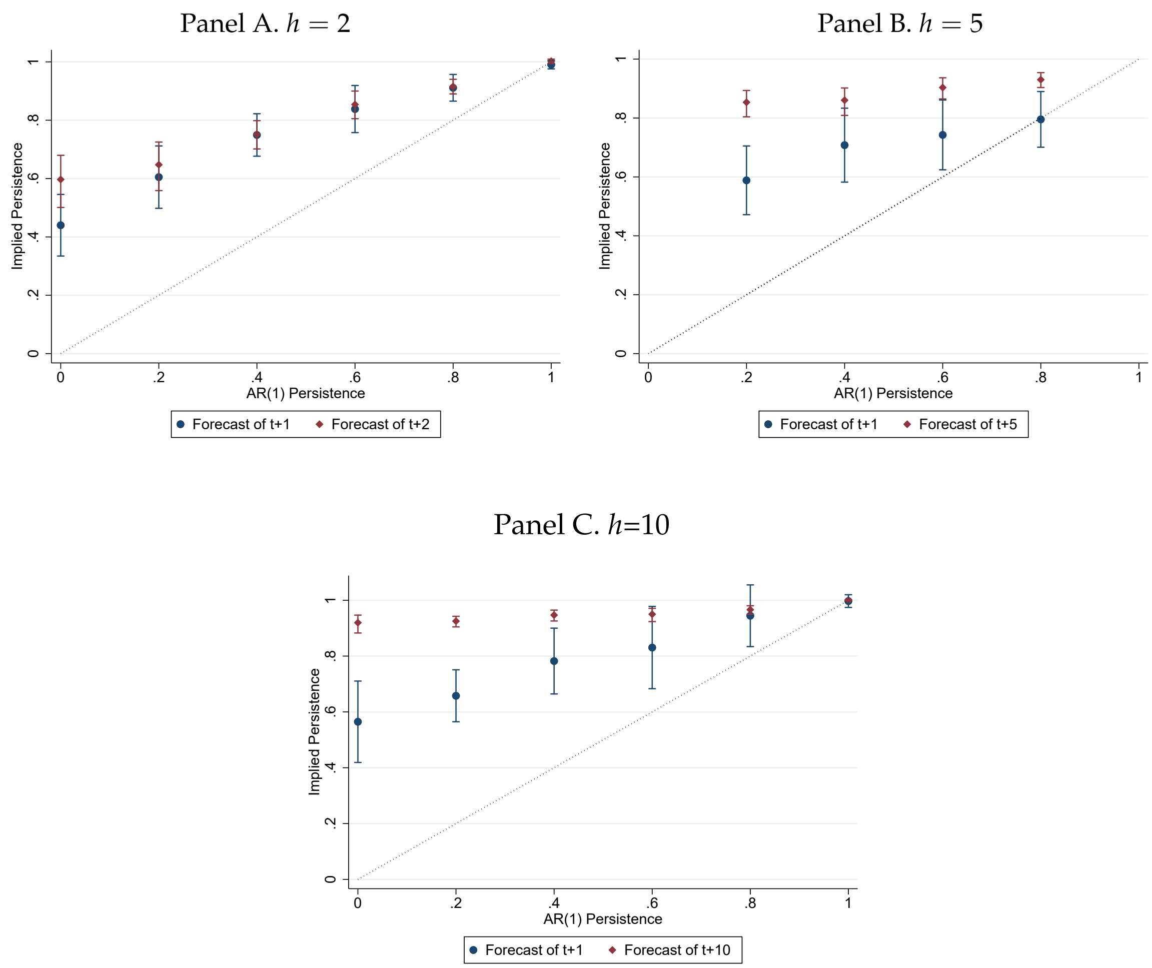
We next examine how overreaction varies across different forecast horizons, by comparing forecasts for x_{t+1} and forecasts for x_{t+2}, x_{t+5}, and x_{t+10}. Recent research using survey data suggests that overreaction appears more pronounced for forecasts of longer-horizon outcomes. For instance, using the error-revision regression, Bordalo et al. (2019) find a negative and significant coefficient for equity analysts' forecasts of long-term earnings growth (which points to overreaction), while Bouchaud et al. (2019) document a positive error-revision coefficient for analysts' forecasts of short-term earnings (which points to underreaction). Based on professional forecasters' predictions of interest rates, several studies also show that the error-revision coefficient is negative and significant for long-term interest rates, but not for short-term interest rates (Bordalo et al., 2020c; Wang, 2020; d'Arienzo, 2020).

We compare the degree of overreaction for long-horizon versus short-horizon forecasts following the structure in Giglio and Kelly (2018). They show that affine asset pricing models using a given level of process persistence cannot simultaneously account for prices of long-maturity and short-maturity claims. We ask if the level of implied persistence (ρ_h^s) differs for long-term and short-term forecasts in our data.

Figure III shows the results. Panels A, B, and C report the values of the implied persis-

Figure III: Implied Persistence for Short-Term and Long-Term Forecasts

This figure shows the implied persistence ρ_h^s as a function of the actual persistence ρ. The implied persistence ρ_h^s is obtained by regressing $F_t x_{t+h}$ on x_t and taking the $1/h$th power of the coefficient. The vertical bars indicate the 95% confidence interval. Panels A, B, and C show results for $h = 2$, $h = 5$, $h = 10$, respectively. The data for Panels A, B, and C come from Experiment 1 (where participants forecast x_{t+1} and x_{t+2}), 2 (where participants forecast x_{t+1} and x_{t+5}), and 4 (where participants forecast x_{t+1} and x_{t+10}) respectively. The two sets of dots (short-horizon and longer-horizon forecasts) for each panel come from forecasts made by the same participants. The dotted line is the 45-degree line. There are no results for $\rho = 0$ and $\rho = 1$, because the experiment initially designed to collect data on $h = 5$ (Experiment 2) did not include these conditions.



tence ρ_h^s for $h = 2$, $h = 5$ (for which we only have conditions with ρ between 0.2 and 0.8), and $h = 10$, respectively. In each panel, the longer-horizon forecasts and the short-horizon forecasts (for x_{t+1}) come from the same participants for the same AR(1) process. In all panels, we see a substantial degree of overreaction. Moreover, the degree of overreaction (as reflected by the gap between the implied persistence ρ_h^s and ρ) is even higher when h is larger. The difference between the implied persistence for short-horizon and longer-horizon forecasts is especially pronounced for very large h and for transitory processes (significant at 5% as the confidence intervals show). The results also suggest that forecasters using a single incorrect persistence parameter (e.g., Gabaix, 2018) cannot fully explain the empirical evidence: this bias alone cannot simultaneously square with both short-term and long-term forecasts made by the same forecasters.

4.2 Testing Models of Expectations

We now use the data from our experiments and the key facts above to analyze commonly-used models of expectations formation.

A. Models of Expectations

We begin by laying out the standard formulations of these expectations models.

Backward-Looking Models

We start with older "backward-looking" models, which specify fixed forecasting rules based on past data and do not incorporate properties of the process (e.g., they are not a function of ρ). The term structure of expectations in these models is not well-defined, so we focus on one-period-ahead forecasts.

1. Adaptive expectations

Adaptive expectations have been used since at least the work of Cagan (1956) on inflation and Nerlove (1958) on cobweb dynamics. The standard specification is:

$$F_t x_{t+1} = \delta x_t + (1 - \delta) F_{t-1} x_t. \tag{4.3}$$

2. Extrapolative expectations

Extrapolative expectations have been used since at least Metzler (1941), and are sometimes used in studies of financial markets (Barberis, Greenwood, Jin and Shleifer, 2015; Hirshleifer, Li and Yu, 2015). One way to specify extrapolation is:

$$F_t x_{t+1} = x_t + \phi(x_t - x_{t-1}). \tag{4.4}$$

That is, expectations are influenced by the current outcome and the recent trend, and $\phi > 0$ captures the degree of extrapolation.

Forward-Looking Models

We now proceed to "forward-looking" models, where forecasters do incorporate features of the true process. In these models, which contain rational expectations, the term structure of expectations is well-defined.

3. Full information rational expectations

Full information rational expectations (FIRE) is the benchmark specification in economic modeling. Decision makers know the true DGP and its parameters, and make statistically optimal forecasts accordingly:

$$F_t x_{t+h} = E_t x_{t+h} = \rho^h x_t. \tag{4.5}$$

As explained in Section 4.1, in our data in-sample least square learning is very close to FIRE, so we use FIRE as the benchmark.

4. Noisy information/sticky expectations

Noisy information models assume that forecasters do not observe the true underlying process, but only noisy signals of it (e.g., Woodford, 2003). In our setting where actual realizations are displayed directly, such frictions may correspond to noisy perception. These models typically have the following recursive formulation:

$$F_t x_{t+h} = (1-\lambda) E_t x_{t+h} + \lambda F_{t-1} x_{t+h} + \epsilon_{it,h}, \tag{4.6}$$

where $E_t x_{t+h}$ is FIRE, $\lambda \in [0, 1]$ depends on the signal's noisiness, and $\epsilon_{it,h}$ comes from the noise too. As a reduced form formulation, Equation (4.6) can also represent anchoring on past forecasts, and Bouchaud et al. (2019) use it to model forecasts of equity analysts.

5. Diagnostic expectations

Diagnostic expectations are introduced by Bordalo, Gennaioli and Shleifer (2018) to capture overreaction in expectations driven by the representativeness heuristic (Kahneman and

Tversky, 1972). The baseline specification is:

$$F_t x_{t+h} = E_t x_{t+h} + \theta(E_t x_{t+h} - E_{t-1} x_{t+h}). \tag{4.7}$$

That is, the subjective expectation is the rational expectation plus the surprise (measured as the change in rational expectations from the past period) weighted by θ, which indexes the severity of the bias. Under diagnostic expectations, subjective beliefs adjust to the true process and incorporate features of rational expectations ("kernel of truth"), but overreact to the latest surprise by degree θ.

6. Constant gain learning

We also test constant gain learning about x_t, which implies least square learning with weights that decrease for observations further in the past. We use the regression specification:

$$F_t x_{t+h} = \widehat{E}_t^m x_{t+h} = \widehat{a_{h,t}} + \widehat{b_{h,t}} x_t, \tag{4.8}$$

where $\widehat{a_{h,t}}, \widehat{b_{h,t}}$ are obtained through a rolling regression with all data available until t. The difference with the standard least square learning specification is that this regression uses decreasing weights (i.e., older observations receive a lower weight) to reflect imperfect retention of past information. Specifically, in period t, for all past observations $s \leq t$, we use exponentially decreasing weights: $w_t^s = \frac{1}{\kappa^{(t-s)}}$. These weights correspond to constant gain learning in recursive least squares formulations (Malmendier and Nagel, 2016).

Other Models

In addition to the models listed above, there are several other models to consider. We do not estimate these models formally because their features are qualitatively different from the results observed in our data, by design or by outcome, as we explain below.

First, an intuitive model of overreaction is described in Gabaix (2018). Specifically, the forecaster faces a range of possible processes with varying degrees of persistence. To limit computational cost, the boundedly rational forecaster uses a persistence parameter $\hat{\rho}$ that is anchored to a default level of persistence ρ^d: $\hat{\rho} = m\rho_i + (1-m)\rho^d$. In such a setting, forecasters would overreact to processes that are less persistent than the default level ρ^d, and underreact to processes that are more persistent than the default level. One limitation of this approach is such a bias alone cannot account for results across different forecast horizons. As Figure III shows, a given level of incorrect persistence cannot simultaneously square with both short-term and long-term forecasts made by the same forecasters. Indeed, if a single incorrect ρ was the only bias, then overreaction would dissipate for long-term forecasts (e.g., forecast of x_{t+10}), which is not the case in the data.

Second, several papers investigate belief formation with model misspecification. For instance, in natural expectations (Fuster, Laibson and Mendel, 2010), the key observation is that forecasters can have difficulty differentiating processes with hump-shaped dynamics (such as AR(2) or ARMA(p,q)) from simpler AR(1) processes in finite samples.[19] Other models analyze subjective beliefs about regime shifts (Barberis, Shleifer and Vishny, 1998; Bloomfield and Hales, 2002; Massey and Wu, 2005). As explained in Section 4.1, in Experiment 3 among MIT EECS students, we explicitly describe the linear AR(1) process to half of the participants. We do not find that the information of a linear AR(1) prior affects the results. Indeed, our findings highlight that systematic biases in expectations can be significant even in linear stationary environments.

Relatedly, building on Rabin (2002), Rabin and Vayanos (2010) also formulate a model based on beliefs about misspecified DGP. Proposition 6 in their paper states that the forecast should have a negative loading on the most recent observation (x_{t-1}), whereas this loading is strongly positive in our data.

B. Estimating Models of Expectations

We now estimate the models described above on one-period-ahead forecast data (i.e., with $h = 1$). We pool data from all conditions of Experiment 1 (i.e., with $\rho \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$). All models except FIRE (which has no parameter) and constant gain learning (whose parameter lies in the decreasing weights) can be simply estimated using constrained least squares. We cluster standard errors at the individual level. The constant gain learning model is estimated by minimizing, over the decay parameter, the mean squared deviation between

[19] Fuster, Laibson and Mendel (2010) formulate an "intuitive model" $F_t x_{t+1} = x_t + \phi(x_t - x_{t-1}) + \epsilon_{t+1}$, when the true DGP is an AR(2) $x_{t+1} = \alpha x_t + \beta x_{t-1} + \eta_{t+1}$, and $\phi = (\alpha - \beta - 1)/2$. We could test this model in our data, where $\alpha \geq 0, \beta = 0, \phi < 0$, and the intuitive model has the same functional form as the extrapolative expectation in Equation (4.4) with negative ϕ.

model-generated and observed forecasts. We estimate standard errors for this model by block-bootstrapping at the individual level.

Table A.5 reports the estimated parameters. Each model is described by an equation and a parameter (in bold). The parameter estimate is reported in the third column, along with standard errors in the fourth column. In the fifth column, we report the mean squared error of each model, as a fraction of the sample variance of forecast. Since forecasts in the $\rho = 1$ condition are mechanically more variable than forecasts in the $\rho = 0$ condition, we compute one such ratio per level of ρ, and then compute the average ratio across values of ρ.

Several patterns emerge from the model estimation. First, consistent with findings in Section 4.1, rational expectations are rejected. Indeed, rational expectations are nested in all three forward-looking non-RE models, and the coefficient related to deviations from rational expectations is always significant at 1%. In line with such deviations being important, FIRE has the lowest explanatory power of forecast data.

Second, most models point to strong signs of overreaction. The adaptive model features overreaction through the high loading on the past realization x_t (0.83). The backward-looking extrapolative model has a negative coefficient on the slope $(x_t - x_{t-1})$, but this again reflects that most overreaction is built into the coefficient on the past realization x_t, which is fixed at one by definition. The diagnostic expectations model has $\theta = 0.34$, which indicates strong overreaction (forecasts react 34% “too much" to the last innovation).[20] The constant gain learning model features a significant decay in the weight of past observations, a loss of 6% per period (i.e., it takes about 12 periods to divide the weight by 2), rejecting the equal weights in benchmark least square learning. Last, the sticky/noisy expectations model is the only one that does not feature overreaction. The coefficient on previous forecasts $(F_{t-1}x_{+1})$ is 0.14, similar to earlier analyses on earnings forecasts by stock analysts (Bouchaud et al., 2019). This finding suggests that there is some anchoring on the level of past forecasts, in addition to overreaction to the recent realization.

[20]The θ estimate is slightly lower than the typical estimate in Bordalo et al. (2020c) using macro survey data (which find θ of around 0.5) and in Bordalo, Gennaioli and Shleifer (2018) and Bordalo et al. (2019) using analyst forecasts of credit spreads and long-term EPS growth (which find θ of around 1).

C. Comparing Model Predictions and Empirical Results

We now investigate how overreaction varies with process persistence and forecast horizon in the estimated models, and compare model predictions with our empirical findings.

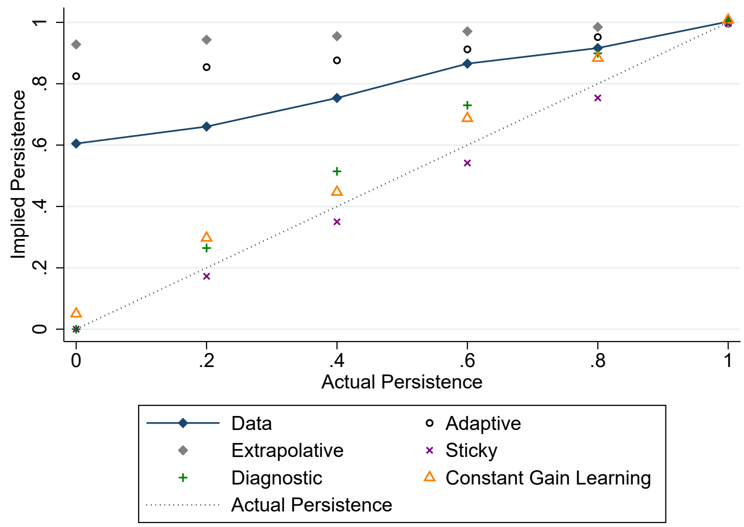
Overreaction and Process Persistence. We start with the evidence on overreaction and process persistence (using one-period-ahead forecasts). In Figure IV, we compute the implied persistence based on the five models estimated above. Specifically, for each model m and for each observation in our data, we compute the model-based forecast $\widehat{F_t^m x_{t+1}}$ using the parameters in Table A.5. We then group observations per level of $\rho \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. For each level, we regress the model-based forecast $\widehat{F_t^m x_{t+1}}$ on x_t to obtain the implied persistence according to the model.

In Figure IV, the solid line represents the implied persistence based on actual forecasts (same as Figure II, Panel B). The dots represent the implied persistence based on the models. In all models, the implied persistence is an increasing function of ρ, and is close to one for random walks (as in rational expectations). However, the commonly-used models do not match the transitory processes very well. The backward-looking expectations models generate too much overreaction for transitory processes, whereas most of the forward-looking models do not generate enough overreaction. Diagnostic expectations and sticky expectations generate no overreaction for i.i.d. processes (e.g., the rational surprise in the benchmark formulation of diagnostic expectations is zero in this case). The constant gain learning model does slightly better: by giving larger weights to recent observations, the model generates some excess sensitivity to recent realizations. Nonetheless, the weights on past observations, as fitted on forecast data, do not decrease fast enough.

To connect with results in field data and for completeness, we also report in Figure A.8 the error-revision coefficients based on the models. Again, the solid line represents experimental data (same as Figure A.8, Panel A) and the dots represent predictions from estimated models. Here we omit the adaptive and extrapolative models, because they do not impose an obvious structure on the two-period-ahead forecasts $F_t x_{t+2}$, which are needed to compute revisions. The conclusions are similar to those in Figure IV. For transitory processes, diagnostic and sticky expectations tend to lead to error-revision coefficients that are too high.

Figure IV: Forecast-Implied Persistence: Data vs Models

For each model m, we compute the model-based forecast $\widehat{F_t^m x_{t+1}}$ for each observation in our data. We use the model parameters reported in Table A.5. We then group observations per level of actual persistence $\rho \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. For each level of ρ, we regress the model-based forecast $\widehat{F_t^m x_{t+1}}$ on lagged realization x_t. The dots report this regression coefficient, which is the forecast implied persistence according to model m for a given level of ρ. The solid line corresponds to the forecast implied persistence in the data, also shown in Figure II, Panel B.



Constant gain learning, on the contrary, generates a coefficient that is too negative.[21]

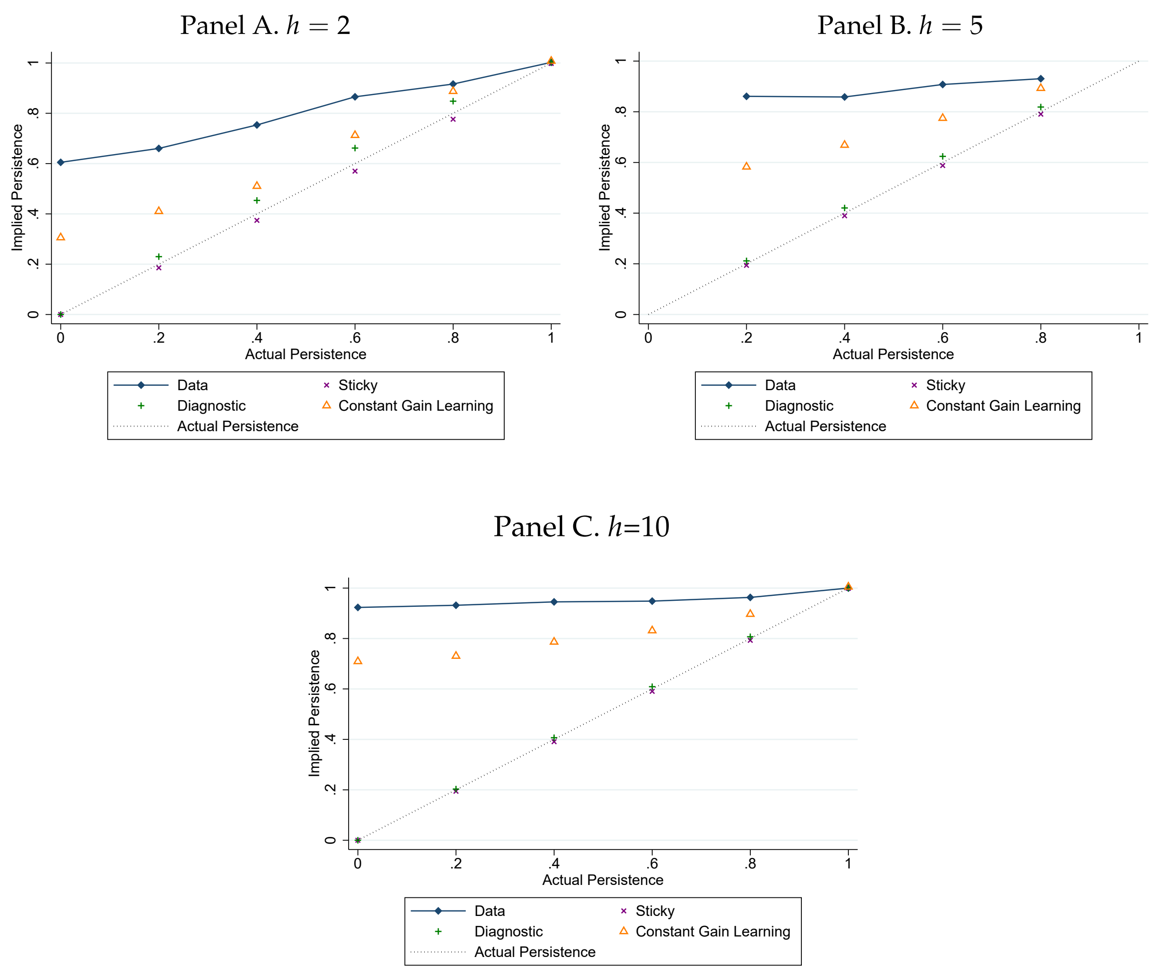
Overreaction and Forecast Horizon. We now compare the data and model predictions for longer-horizon forecasts. We focus on the forward-looking models as the backward-looking models do not provide a clear term structure of forecasts for multiple horizons. In particular, we fit all models using $h = 1$ and use the same parameters to generate model predictions for $h = 2, 5, 10$. We find that the implied persistence according to standard models tends to be too low, especially when the forecast horizon is longer; the exception is constant gain learning, which produces a closer fit when the horizon is longer.

Overall, the empirical findings suggest that forecasts in the data do adapt to the setting (with lower implied persistence when the true persistence ρ is smaller). Consequently, the backward-looking models (which specify a fixed dependence of forecasts on past realizations) generally do not perform well in capturing the degree of overreaction across different

[21] This is in fact a mechanical effect of the error-revision coefficient, which divides by the variance of forecast revision. In the constant gain learning model, forecast revisions tend to be very small for low values of ρs (they are close to zero), which blows up the absolute value of the error-revision coefficient. The implied persistence measure in Figure IV is immune to this problem.

Figure V: Forecast-Implied Persistence: Longer Horizon Forecasts

This figure shows the implied persistence ρ_h^s as a function of the actual persistence ρ. The subjective persistence ρ_h^s is obtained by regressing $F_t x_{t+h}$ on x_t and taking the $1/h$th power of the coefficient. Panels A, B, and C show results for $h = 2$, $h = 5$, $h = 10$, respectively. The data is the same as those used in Figure III. The dotted line is the 45-degree line.



levels of process persistence. Meanwhile, despite the partial adaptation of forecasts to the actual persistence, overreaction is most significant when the process is transitory. This result echoes strong extrapolative beliefs observed in other settings with transitory processes, including experimental studies (Frydman and Nave, 2016) and survey data on stock returns (Greenwood and Shleifer, 2014). The baseline specifications of the forward-looking models generally do not allow transitory shocks to generate much overreaction. This is because these models focus on using the recent shock ϵ_t to infer future shocks ϵ_{t+h}, which tends to produce an inferred value ($F_t[\epsilon_{t+h}|\epsilon_t]$) that scales with the persistence ρ.

The empirical evidence points to the following insight: a parsimonious way for models to account for the empirical findings is allowing recent observations to influence the assessment of the *long-run mean* of the process. The benefit of modeling imperfect inference about the long-run mean is two-fold. First, this approach naturally implies that recent observations can lead to biased forecasts of future outcomes even when the process is i.i.d. Second, this approach also naturally leads to greater overreaction of longer horizon forecasts, which are more sensitive to biases regarding the long-run mean. For the models discussed above, for example, one can modify constant gain learning to focus on the mean instead of x_{t+h}, as in Nagel and Xu (2019). Similarly, da Silveira, Sung and Woodford (2020) provide a model where the noisiness of the memory state leads the agent to put a large weight on the recent observation when inferring the mean. The biases about the long-run mean can be generalized to other models as well. For the diagnostic expectations model, the agent can start at each point in time with a default prior regarding μ, with a higher realization of x_t diagnostic of a higher μ.[22]

In the next section, we present a simple and tractable model that builds on this insight of biases about the mean. The model delivers three key features. First, it microfounds the overweighting of recent observations as a result of imperfect information processing (Sections 5.1 and 5.2). Second, it can be easily estimated to evaluate empirical performance and readily generalized to different processes and settings (Section 5.4 and Appendix C). Third, it can accommodate both overreaction and underreaction (Appendix D). The model generates further predictions regarding the effect of influencing what information is easier to process, which we test in Section 5.5.

5 Model

In this section, we provide a simple model that captures the disproportionate influence of recent observations on expectations, which operates through judgment about the mean of the process. We set up and solve the model in Sections 5.1 and 5.2. We show the basic comparative statics regarding the degree of overreaction in Section 5.3 (and discuss extensions for

[22]This approach is similar to that of Bordalo et al. (2019), where agents partially conflate transitory noise with a more persistent underlying latent process. (In our setting, however, there is no unobservable state.)

underreaction). Finally, we assess the model fit in Section 5.4, and test additional predictions in Section 5.5.

5.1 Setup

Environment. Time is discrete and is indexed by $t \in \{0, 1, 2, \ldots\}$. There is an agent who forecasts future realizations of an exogenous stochastic process $\{x_t : t \geq 0\}$ at horizon h. The process is AR(1) with mean μ and persistence ρ:[23]

$$x_t = (1-\rho)\mu + \rho x_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2). \tag{5.1}$$

The agent's payoff at any given time t depends on the accuracy of these forecasts and is given by: $-(F_t x_{t+h} - x_{t+h})^2$, where $F_t x_{t+h}$ is the agent's time t forecast of x's realization h periods ahead and x_{t+h} is the *ex post* realization of the variable at $t+h$.[24]

Information Processing. We assume that the agent is uncertain about the long-run mean (μ) but can process information to form an assessment of its value in order to forecast x_{t+h}. We model this using information processing where more recent observations are less costly to process than others. The simplest way to obtain such a cost structure is to assume that the agent processes the most recent observation x_t for free, and processes further information at a cost.[25] Formally, after observing x_t, the agent automatically forms the initial prior regarding $\mu \sim N\left(x_t, \underline{\tau}^{-1}\right)$. The agent then decides whether to process more information to update this prior. Letting S_t denote the set of all processed information (which includes x_t), we assume the cost of processing S_t is increasing and convex in the amount of information

[23]The model can be extended to a general Gaussian ARMA processes and the qualitative conclusions of the model are unchanged. See Appendix C for details.

[24]Note that x_{t+h} is not fully known at time t and only realized h periods after the forecast is made. Nonetheless, at time t, the agent knows that the payoff is determined by the realization of the process at $t+h$. This is similar to the score function in the experiment. A minor difference is that the score function in the experiment does not have an exact quadratic form to ensure that payments in the experiment are always non-negative (as discussed in Section 3.2). We use this standard quadratic form for simplicity of modeling, so we can derive closed-form solutions.

[25]More generally, one can allow for more observations to be processed for free or all observations to be costly to process. Our model predictions are largely unchanged as long as information processing is relatively cheaper for recent observations than for others. In line with our main assumption, Table A.6 shows that x_t has a disproportionate impact on both the forecast and the deviation from the rational benchmark in our data.

processed conditional on x_t:

$$C_t(S_t) \equiv \omega \frac{\exp\left(\gamma \cdot \mathbb{I}(S_t, \mu | x_t)\right) - 1}{\gamma},$$

where $\omega \geq 0$ and $\gamma \geq 0$ are the scale and convexity parameters, and $\mathbb{I}(S_t, \mu|x_t)$ is Shannon's mutual information function, which measures the amount of information utilized by the agent after observing x_t.[26] We use the term "on top of mind" to refer to the set of information actively utilized by the agent, S_t. In the extreme case when processing any additional information is costless, our model nests the frictionless rational benchmark. However, with costly information processing, only a subset of data will be on top of mind.

Psychological Foundations. Our key assumption for information processing is that recent information is easier to utilize. First, this assumption is related to a series of research in psychology on working memory, which emphasizes that some information is more actively utilized than others. This notion has been referred to as heightened activation, increased accessibility, conscious awareness, or focus of attention (Baddeley and Hitch, 1993; Cowan, 1998, 2017a), and is connected more broadly to costly information processing in a variety of settings (Spillers, Brewer and Unsworth, 2012).[27] We use the term "on top of mind" to refer to the set of information actively utilized, which corresponds to the set S_t defined above.

Second, this literature also highlights how the most recently processed items tend to be especially on top of mind and most strongly expressed in behavior, which provides a foundation for recency effects. Baddeley and Hitch (1993) and Baddeley (2007) present a comprehensive summary of findings about recency effects in psychology and argue that a theory based on higher activation of more recent observations is consistent with the evidence. Such recency effects can be present even in settings that do not explicitly involve recall.[28] Even though a person may see a large number of observations in her environment

[26] This functional form embeds two useful cases. First, it becomes linear in $\mathbb{I}(S_t, \mu|x_t)$ when $\gamma \to 0$, which is the classic formulation in rational inattention (Sims, 2003). Second, in case of Gaussian beliefs with $\gamma > 1$, the cost is equivalent to choosing the precision of beliefs about μ (see Appendix B.2 for formal derivations).

[27] The recent survey paper by Cowan (2017a) explains the concept of working memory as "the ensemble of components of the mind that hold information temporarily in a heightened state of availability for use in ongoing information processing." Cowan (2017b) conveys the idea through its title, "Working Memory: The Information You Are Now Thinking of."

[28] The working memory literature emphasizes that heightened activation is not necessarily about recall but rather about focusing on a subset of the available information in processing. See, e.g., Unsworth and Spillers

(e.g., many data points on the screen), not all information is necessarily actively processed. As Barrett, Tugade and Engle (2004) write: "any environment contains an array of stimuli ... but stronger representations will laterally inhibit weaker ones, and the strongest will be expressed in behavior." Baddeley and Hitch (1993) write: the "presentation and processing of an item results in the activation of its node... [and] the recency effect occurs because recently activated nodes are easy to reactivate."[29] Our model builds on this notion of recency whereby the most recent observation is less costly to process.

Finally, the assumption that the individual starts from a default driven by what is immediately accessible ("System 1") and further adjusts beliefs by effortful processing ("System 2") also connects to research in psychology and economics on dual process models, which can be microfounded by working memory as well (Barrett, Tugade and Engle, 2004). See Evans (2008) for a summary of the many types of dual process models in the psychology literature and Ilut and Valchev (2020) for an application of dual process in economics.

Taken together, our model aims to capture the general insight that recent observations have a disproportionate influence on beliefs. This can reflect costly information processing as discussed above, but more generally can also arise from other psychological and institutional mechanisms that limit the usage of past data and generate recency effects.

Forecasts. Given her estimate of the long-run mean conditional on S_t, $\mathbb{E}[\mu|S_t]$, the agent's forecast of $F_t x_{t+h}$ is:

$$F_t x_{t+h} = \rho^h x_t + (1-\rho^h)\mathbb{E}[\mu|S_t] = \underbrace{E_t x_{t+h}}_{\text{rational forecast}} + \underbrace{(1-\rho^h)(\mathbb{E}[\mu|S_t]-\mu)}_{\text{forecast error of long-run mean}}. \tag{5.2}$$

We assume that the agent uses the correct ρ for simplicity. As discussed earlier, modeling frictions in beliefs about the long-run mean μ is the most parsimonious way to capture how overreaction varies with the process persistence and forecast horizon, whereas biases in ρ

(2010) for a discussion. Cowan (2017a) also explains that the term "working memory" originated from the use in computer science (where it refers to holding information for active processing). Barrett, Tugade and Engle (2004) discuss several related concepts about the processing component of working memory: "what Baddeley and Hitch (1974) called the central executive, what Norman and Shallice (1986) called the supervisory attention system (SAS), and what Posner and DiGirolamo (2000) called executive control".

[29] For a broader discussion of recency effects in psychology and economics, see also Kahana (2012) and Bordalo, Gennaioli and Shleifer (2020b).

by itself is not sufficient (as shown by the results across different forecast horizons). Overall, we do not rule out that forecasters may also use an incorrect ρ. Nonetheless, we find that modeling biases about the mean μ is the most concise approach to capture the empirical evidence that overreaction is stronger both when the process is less persistent and when the forecast horizon is longer.

5.2 Model Solution

Given the primitives of the problem at time t, the agent solves:

$$\min_{S_t} \mathbb{E}\Big[\min_{F_t x_{t+h}} \mathbb{E}\left[(F_t x_{t+h} - x_{t+h})^2 | S_t\right] + C_t(S_t)\Big] \tag{5.3}$$

$$s.t. \quad \underbrace{\{x_t\}}_{\text{recent observation}} \subseteq \underbrace{S_t}_{\text{actively utilized information}} \subseteq \underbrace{\mathcal{A}_t}_{\text{total information}}$$

where $\mathcal{A}_t$ is the largest possible set of information that is available for processing given the set of available observations $x^t \equiv \{x_\tau\}_{\tau \leq t}$.[30] In Appendix B.2, we show that the above problem can be simplified to choosing the optimal precision of the long-run mean estimate:

Lemma 1. *The agent's problem in Equation (5.3) is equivalent to:*

$$\min_{\underline{\tau} \leq \tau \leq \bar{\tau}_t} \Bigg\{ \underbrace{\frac{(1-\rho^h)^2}{\tau}}_{\textit{benefit of precision}} + \underbrace{\omega \frac{\left(\frac{\tau}{\underline{\tau}}\right)^\gamma - 1}{\gamma}}_{\textit{cost of precision}} \Bigg\}, \tag{5.4}$$

where $\tau \equiv var(\mu | S_t)^{-1}$ is the precision of the agent's posterior belief about the long-run mean, and $\bar{\tau}_t$ is the maximum precision obtainable given the full information set $\mathcal{A}_t$.

For the remainder of the paper, we assume that the constraint $\tau^* \leq \bar{\tau}_t$ does not bind, which occurs when $\mathcal{A}_t$ is sufficiently large. As formally shown in Appendix B.3, it is straightforward to derive the optimal posterior precision of the long-run mean, $\tau^* = \text{var}(\mu | S^t)^{-1}$, as $\tau^* = \underline{\tau} \max\left\{1, \left(\frac{(1-\rho^h)^2}{\omega \underline{\tau}}\right)^{\frac{1}{1+\gamma}}\right\}$, with the agent's forecast error of the long-run mean given

[30]Formally, $\mathcal{A}_t \equiv \{s | \mathbb{I}(s, \mu | x^t) = 0\}$, meaning that no available signal should contain further information about the long-run than what is revealed by the history of available observations.

by:

$$\mathbb{E}[\mu|S_t] - \mu = \frac{\underline{\tau}}{\tau^*}(x_t - \mu) + \text{noise}. \tag{5.5}$$

Equation (5.5) shows that the agent's forecast error of the long-run mean is anchored towards x_t, which follows from the assumption that x_t is easier to process than other information. The dependence of the agent's perception of the long-run mean on the most recent observation is the key force that drives the overreaction in forecasts. By applying Equation (5.5) to Equation (5.2), we obtain the following proposition, which presents the behavior of forecasts under the optimal information processing.

Proposition 1. Forecasts display systematic overreaction relative to the rational benchmark, with

$$F_t x_{t+h} = \underbrace{E_t x_{t+h}}_{\text{rational forecast}} + \underbrace{(1-\rho^h)\min\left\{1, \left(\frac{\omega\underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}\right\} x_t}_{\text{overreaction } (\equiv \Delta)} + \underbrace{u_t}_{\text{noise}}. \tag{5.6}$$

Proof. See Appendix B.3. □

Proposition 1 shows that in presence of costly information processing, forecasts are *more sensitive* to the most recent observation relative to the rational forecasts formed under full information. We denote this excess sensitivity by the term Δ, which captures the degree of overreaction in forecasts relative to rational forecasts.

5.3 Comparative Statics

We now illustrate the implications of our model for the empirical evidence on overreaction. As Proposition 1 shows, forecasts display overreaction in our model since the agent overweighs the most recent observation x_t when inferring the long-run mean μ.[31] Furthermore, as the agent conflates some part of the transitory shock as permanent, we also obtain predictions about how the degree of overreaction varies across different settings. The key intuition

[31] This is a fundamental difference between our model and models of sticky information (which may use similar modeling techniques). In sticky information models, agents have full access to past information, but some may not have access to the most recent observation. Accordingly, forecasts can exhibit underreaction (since they rely more on past information rather on the recent observation). In contrast, in our model, agents are fully aware of the most recent observation and they have to decide the extent to process past data, which results in overreaction (since forecasts rely more on the recent observation than on past information).

is as follows. First, overreaction is stronger when the most recent shock is less predictive of future outcomes. Accordingly, our model naturally generates greater overreaction for less persistent processes and for longer-horizon forecasts. Second, as Equation (5.6) shows, forecasts' response to x_t in our model also adapts partially to the true process (in line with what we see in the data). In other words, costly information processing allows the agent to moderate the influence of x_t on her forecast. Therefore, the implied persistence is lower when actual ρ is smaller, instead of being a constant, but the adjustment is imperfect.

The following proposition summarizes the combined effect of these two forces and provides comparative statistics of overreaction with respect to the parameters of the model.

Proposition 2. Consider the excess sensitivity of forecasts to x_t measured by $\Delta = \rho_{s,h} - \rho^h$ defined in Equation (5.6):

1. $\Delta \geq 0$ with $\Delta = 0$ if and only if, either $\rho = 1$, or information processing is frictionless ($\omega = 0$) and past information available to the forecaster is infinite.
2. Δ is increasing in $\underline{\tau}$ and ω.
3. Δ is decreasing in ρ^h (and hence decreasing in ρ and increasing in h) if the cost function is weakly convex in τ, which is true if and only if $\gamma \geq 1$.
4. Denote $\zeta(\rho, h) \equiv \rho_{s,h}/\rho^h$. Then, $\zeta(\rho, h) \geq 1$. Like Δ, $\zeta(\rho, h)$ is decreasing in ρ and increasing in h for all values of ρ and h, if and only if $\gamma \geq 1$.

Proof. See Appendix B.4. □

Proposition 2 implies that the excess sensitivity of forecasts to the most recent observation (as measured by Δ or $\zeta(\rho, h)$), is decreasing in the persistence ρ and increasing in the horizon h, as long as the cost of information processing is weakly convex in the precision τ. Moreover, this insight carries over to the gap between the implied persistence ρ_h^s and the true persistence ρ, especially for less persistent processes.[32] We derive this formally in the following corollary and show the numerical results in Figure VI.

[32]More precisely, we show in Appendix B.5 that $\Delta\rho_h$ is monotonically decreasing in ρ^h for $\rho^h \leq \lambda$, a positive constant independent of ρ and h which depends on γ and $\omega\underline{\tau}$. In practice, as shown in Figure VI, we find $\Delta\rho_h$ is increasing in the range of persistence and horizon covered in our experiment for our calibrated parameters.

Corollary 1. *Consider the measure of the implied persistence per period relative to the actual persistence* $\Delta\rho_h \equiv \rho_h^s - \rho$*, and also assume* $\gamma \geq 1$*. Then,* $\Delta\rho_h \geq 0$ *and decreasing in* ρ*. Moreover,* $\Delta\rho_h$ *is increasing in* h *for* h *sufficiently large: holding fixed* ρ*, there exists an* h^* *such that* $\Delta\rho_h$ *is increasing in* h *for* $h \geq h^*$*.*

Proof. See Appendix B.5. □

In sum, Proposition 2, along with Corollary 1, delivers two main results of our model. The first result is overreaction, a prediction that is consistent with the evidence presented in Section 4: the gap between the implied persistence and the actual persistence, Δ and $\Delta\rho_h$, is positive (or equivalently, $\zeta(\rho, h)$ is greater than 1). The second result is that overreaction (as measured by Δ, $\Delta\rho_h$, or $\zeta(\rho, h)$) is stronger for less persistent processes and sufficiently long horizons, as we observe in the data (as long as the cost of information utilization is convex in the precision of the agent's forecast).[33]

Finally, the baseline version of the model focuses on overreaction in light of our empirical evidence presented in Section 4. We provide an extension in Internet Appendix Section D that shows how the model can allow for underreaction as well. In particular, if the signals are noisy (Woodford, 2003) or if updating is infrequent (Mankiw and Reis, 2002), then there can be an additional force that push in the direction of underreaction. In this case, the model shows that overreaction will still be relatively more pronounced when the process is less persistent. In our experiment, the signal is clear and infrequent updating is unlikely, so overreaction dominates.

5.4 Model Fit

In addition to the qualitative comparative statics presented above, we now estimate the model using our forecast data to further evaluate its performance. We present results on

[33]Another approach for modeling overreaction is the one in da Silveira, Sung and Woodford (2020) who introduce a dynamic framework where memory is costly and agents optimally choose their memory structure over time. In their model, agents decide what they want to remember in the future before an observation is realized. In our model, the recent observation is the starting point and agents decide to utilize past information after an observation has been realized. While both our model and the model in da Silveira, Sung and Woodford (2020) deliver overreaction in posterior beliefs, the predictions for prior beliefs are different: in our model, the priors are anchored to the present, as the utilization of past data happens after the most recent observation is realized; in da Silveira, Sung and Woodford (2020), in contrast, the priors depend on the memory state which is formed optimally based on the past observations before the most recent observation is realized.

model fit for the case where the cost of information utilization is quadratic ($\gamma = 2$). We also present robustness checks in Section 6.1 where we jointly estimate γ with the other parameters of the model, which produce similar results. As before, the model is estimated by minimizing the mean-squared error (MSE) between the one-period forecast predicted by the model for a given parameter (using the realizations of x_t in the data) and the one-period forecast observed in the data.

First, we show that our model matches the relationship between the implied persistence and the actual persistence found in the data. Figure VI, Panel A, shows the results for the baseline horizon $h = 1$. The solid line represents the implied persistence ρ_1^s in the data, and the red solid circles represent ρ_1^s predicted by our model. We see that the implied persistence ρ_1^s predicted by our model is very similar to that in the data. Note that there is nothing mechanical in this very good fit. The models investigated in Figure IV were fitted the same way as our model, and do not fit the empirical relation between ρ_1^s and ρ.

We also examine model fit for longer-horizon forecasts in Figure VI, Panels B to D. We present the implied persistence (per period) ρ_h^s for horizons $h = 2, 5, 10$. Importantly, we fit our model using one-period-ahead forecasts, so its performance for other forecast horizons is not targeted. We see that our model performs well for all forecast horizons.

Table A.7 further evaluates the model fit by calculating the MSE between ρ_h^s in the model and ρ_h^s in the data, as well as the MSE between $F_t x_{t+h}$ in the model and in the data. We calculate the MSE for our model and the models in Section 4.2. This MSE calculation also confirms that our model performs well.

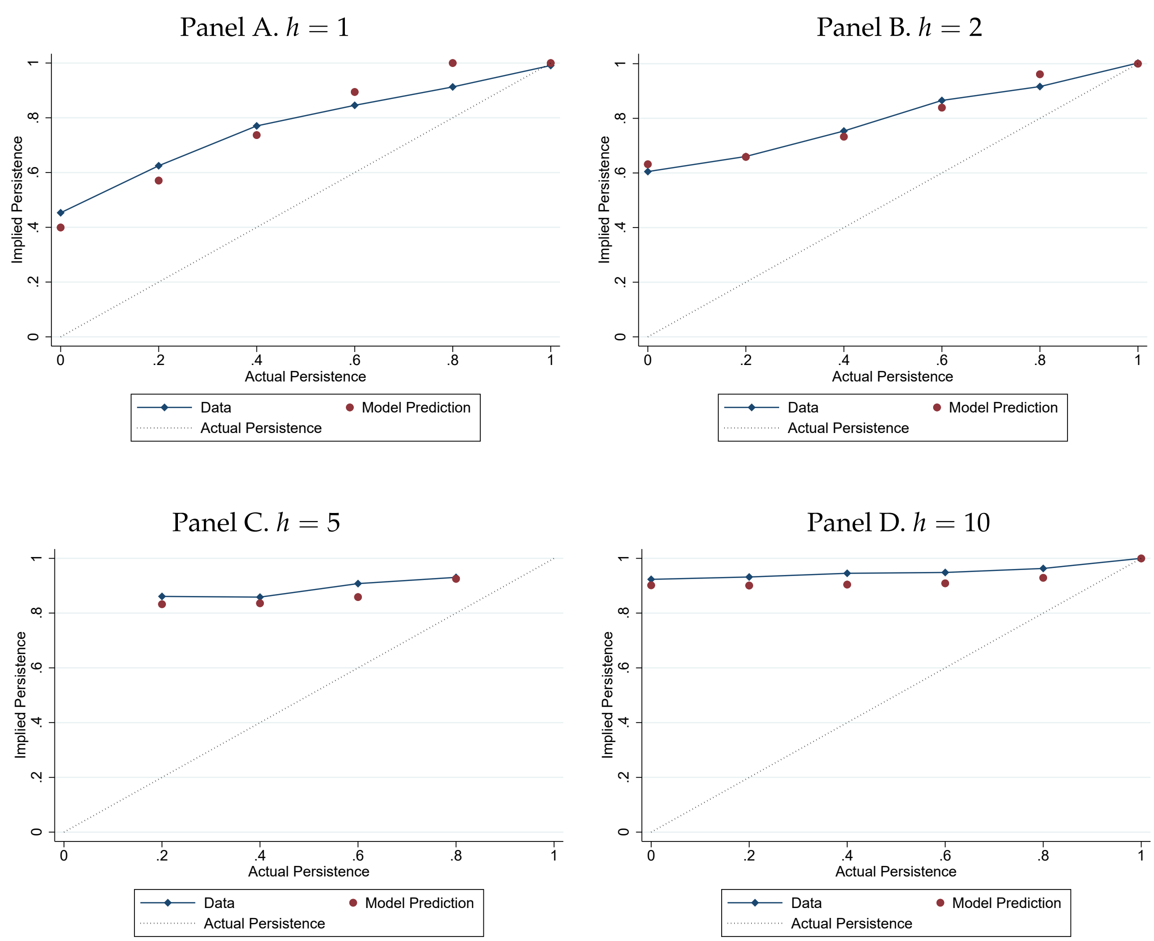
5.5 Changing What's on Top of Mind

Finally, we design additional experiments to investigate the mechanism in our model. A key component of our model is that the recent observation is more on top of mind and more easily available. As a result, forecasters rely too much on it in their judgment of the long-run mean, which leads to overreaction. In this section, we implement additional experiments to test this mechanism directly, by changing the extent to which the last observation is on top of mind.

We design two conditions in which participants are led to rely less on the most recent

Figure VI: Model Fit: Implied Persistence

This figure shows the forecast implied persistence ρ_h^s as a function of the objective persistence ρ. The implied persistence ρ_h^s is obtained by regressing $F_t x_{t+h}$ on x_t and taking the regression coefficient to the $(1/h)$th power. The blue line represents the results in the forecast data. The solid red dot represents ρ_1^s from our model. Results for $h = 1$ and $h = 2$ use forecasts in Experiment 1. Results for $h = 5$ and $h = 10$ use forecasts in Experiment 2 and Experiment 4, respectively. We fit the model parameter using data for $h = 1$ and use the same parameter to generate model predictions for other forecast horizons.



observation. In the first condition ("click x_{t-10}"), we require participants to click on x_{t-10} before making their forecasts in each round. A screenshot of the interface is presented in Panel A of Figure A.9. In the second condition ("red line"), we draw a red line at zero (the actual long-term mean of the process) on the graphical interface for forecasting. A screenshot of the interface is presented in Panel B of Figure A.9. We also include the baseline treatment condition (same design as Experiment 1) for comparison. All conditions contain $\rho \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. Each participant is randomly assigned to a given ρ and a given treatment condition. The data is collected in Experiment 4.

The two additional treatment conditions seek to divert the focus away from the most recent observation, which can reduce its impact on the assessment of the long-run mean. Formally, both treatments can be modeled as bringing in other signals about the long-run mean to top of mind, in addition to the existing default belief $\mu \sim N(x_t, \underline{\tau}^{-1})$. We obtain the following prediction:

Proposition 3. The implied persistence in the new treatment conditions should be less than that in the baseline condition: $\rho^s_{1,I} < \rho^s_1$ for each level of actual ρ (except for $\rho = 1$).

Proof. See Appendix E. $\square$

Intuitively, the provision of the extra signal attenuates the reliance on x_t in forming an assessment about the long-run mean, thereby reducing overreaction. We test Proposition 3 by running standard regressions of forecast error predictability, pooling together the three treatment conditions (the baseline condition and the two additional treatment conditions):

$$x_{it+1} - F_t x_{it+1} = \alpha x_{it} + \beta T_i^{\text{Click } x_{t-10}} \times x_{it} + \gamma T_i^{\text{Red Line}} \times x_{it} + a_i + \epsilon_{it}, \tag{5.7}$$

where $T_i^{\text{Click } x_{t-10}}$ and $T_i^{\text{Red Line}}$ are indicator variables that equal one if individual i is assigned to one of the new treatment conditions, and ϵ_{it} is clustered by forecaster. Since there is strong overreaction in the experiment, we expect $\alpha < 0$. But we expect overreaction to be less pronounced in both conditions where the last observation is less on top of mind, so that $\beta > 0$ and $\gamma > 0$.

Table I shows the results of estimating Equation (5.7). We use forecast errors based on both realizations and full-information rational forecasts. We also look at forecasts for x_{t+1} as well as x_{t+2}. We find that both treatments significantly reduce overreaction, in line with Proposition 3.

6 Robustness and Additional Field Evidence

In Section 6.1, we show the robustness of our model formulations to different functional forms. We also discuss several modeling assumptions. Finally, in addition to analyzing forecast data, we provide supportive evidence from financial markets in Section 6.2.

Table I: Changing What Is on Top of Mind

In this Table, we regress different definitions of the forecast error (realization minus forecast) on the last realization, interacted with two indicator variables that equal one when the participant is allocated to the new treatment conditions. One of these conditions requires participants to click on the point corresponding to x_{t-10} in each round before entering new forecasts. The other one features a red line at $x = 0$. We also include the baseline condition (same design as Experiment 1) for comparison. The data is collected in Experiment 4. Each participant is randomly assigned to a given ρ and a given condition. Regressions include participant fixed effects to control for average optimism. *** indicates a 1% level of significance. Standard errors clustered by participant are presented in parentheses.

	$x_{t+1} - F_t x_{t+1}$	$\rho x_t - F_t x_{t+1}$	$x_{t+2} - F_t x_{t+2}$	$\rho^2 x_t - F_t x_{t+2}$
x_{it}	-0.22***	-0.15***	-0.33***	-0.21***
	(0.029)	(0.030)	(0.031)	(0.034)
$\times$ Click on x_{t-10}	0.15**	0.15***	0.093	0.11*
	(0.058)	(0.055)	(0.067)	(0.059)
$\times$ Red Line at 0	0.14***	0.12**	0.22***	0.21***
	(0.046)	(0.047)	(0.059)	(0.060)
Observations	21,645	21,645	21,090	21,645
R^2	0.22	0.29	0.24	0.32
Participant FE	Y	Y	Y	Y

6.1 Robustness of Model Formulations

We discuss several main assumptions in our baseline model in Section 5.

A. Convexity and General Functional Form

Our benchmark calibration assumes that the cost of information utilization is quadratic ($\gamma = 2$) in the relative precision $\frac{\tau}{\underline{\tau}}$. Here, we examine two alternative ways for calibrating γ and show the robustness of the results. First, we fit our model assuming the cost is linear in the mutual information ($\gamma \mapsto 0$), which is a standard approach in the rational inattention literature (e.g. Sims, 2003). Second, we fully optimize over the convexity parameter γ using a grid-search method. Figure A.10 shows the results. The linear approach does a reasonable job fitting the implied persistence, but overshoots slightly for processes with higher persistence and undershoots slightly for processes with lower persistence. The general γ approach produces very good fit (the optimal value of γ roughly equals 10). Overall, we find that the model has good performance and is not very sensitive to the exact value used for γ.

B. Assumptions on $\underline{\tau}$

Our main model defines $\underline{\tau}$ as the baseline precision the agent has regarding the long-run mean after seeing the most recent observation. For simplicity, we have assumed that $\underline{\tau}$ is fixed across all experiments and across different persistence levels ρ. In the following, we also consider an alternative approach, where we endogenize $\underline{\tau}$. One natural candidate for $\underline{\tau}$ is the inverse of the variance of the stationary distribution for the AR(1) process:

$$\underline{\tau}^{alt} = \frac{1-\rho^2}{\sigma_\epsilon^2}. \tag{6.1}$$

This choice can have a Bayesian interpretation as the posterior variance given x_t, for a Bayesian with an improper uniform prior (or a sequence of priors that become increasingly dispersed). In particular, $\underline{\tau}^{alt}$ is decreasing in ρ: the agent is ex ante more uncertain about the long-run mean when the process in unconditionally more volatile. Figure A.11 shows the fit of the alternative specification and confirms the model performs well in this case too.

C. Incentives and External Validity

A possible question is whether one can test the effect of changing forecasters' incentives, or the trade-off between the cost of information processing and the benefit of obtaining accurate beliefs. While in principle one could design experiments with different incentive schemes, we have refrained from doing so for several reasons. First, to provide a direct test of the role of incentive schemes, we need to *randomly assign* participants to some conditions that pay much more and other conditions that pay much less.[34] This design may be questionable to human subject reviews and may antagonize potential participants when they read disclosures of payments in the consent form. Second, DellaVigna and Pope (2017) also suggest that participants are often not only motivated by monetary incentives.

Another possible question is whether incentives for accuracy in practice could be so large that decision makers will overcome all costs of information utilization. A large literature

[34] For instance, we cannot run the experiment once in a richer country and once in a poorer country with the same dollar payment to create variation in incentives (weaker incentives in the richer country and stronger incentives in the poorer country). The results from the richer country and the poorer country can be different for many other reasons, such as the demographics and quality of participation across countries. Instead, to isolate the treatment effects, we need to start with a given pool of participants and randomly assign them to different conditions in order.

document biases in high-stake settings (Malmendier and Tate, 2005; Pope and Schweitzer, 2011; Ben-David, Graham and Harvey, 2013; Greenwood and Hanson, 2015; Bordalo, Gennaioli, La Porta and Shleifer, 2019), which indicate that frictions may not be fully eradicated in these situations. Furthermore, many decisions are made under time constraints or with a fair bit of human discretion, in which case the frictions represented by our model—namely, certain information is particularly on top of mind—are likely to be present. To assess the external validity of our findings, we also analyze data from financial markets. We present supportive evidence from survey forecasts in Section 2.1 and from asset prices in the next section, which points to variations of overreaction that are in line with our main results.

6.2 Further Results from Financial Markets

Our main analyses of forecast data in previous sections show that overreaction is stronger for more transitory processes and for longer forecast horizons. In the following, we also investigate data on asset prices and explore the implication of our results for financial markets.

For evidence across different horizons, Giglio and Kelly (2018) provide evidence of greater overreaction for longer horizons in asset prices. In particular, they find "excess volatility" for long-maturity claims relative to short-maturity claims. Consistent with our findings in forecast data, Giglio and Kelly (2018) show that a given level of persistence cannot simultaneously account for the behavior of asset prices for long-horizon and short-horizon claims, with "much higher persistence implied from the long end."

For evidence across different levels of process persistence, we analyze stock prices, where a series of papers document the link between investor overreaction and the value premium, namely stocks with high book-to-market ratio have high subsequent returns (e.g. Lakonishok, Shleifer and Vishny (1994), La Porta (1996), Bordalo et al. (2019)). The intuition is that high book-to-market stocks are "cheap" on average because investors overreacted to bad news in the past. Therefore, these stocks are undervalued and their future returns are higher. The reverse holds for low book-to-market stocks. If overreaction is more pronounced for firms with transitory shocks, we would expect this predictability to be stronger for firms with transitory sales processes.

We test this hypothesis using Fama-MacBeth regressions of the following form:

$$r_{it+1} = \alpha + \beta BM_{it} + \gamma\rho_i + \delta BM_{it} \times \rho_i + \epsilon_{it}, \tag{6.2}$$

where r_{it+1} is the annual stock return during fiscal year $t+1$ and BM_{it} is the book-to-market ratio of equity at the end of fiscal year t. The value ρ_i is the persistence of annual sales growth for firm i (i.e., regression coefficient of $\Delta \log sales_{it+1}$ on lagged $\Delta \log sales_{it}$). The mean (median) of ρ_i is 0.16 (0.15) and the inter-quartile range is -0.07 to 0.39. Since we run regressions firm by firm to calculate ρ_i, this estimate may be downward biased if the time series is too short. The median number of observations per firm is 18. We also present robustness checks where we restrict our regressions to firms that have at least 10 observations and the results are essentially unchanged. Overall, we expect that the coefficient of the interaction term $\delta < 0$: firms with less persistent sales processes should have stronger overreaction and therefore a more pronounced value premium.

We report Fama-MacBeth regression results of Equation (6.2) in Table II. Our dataset is a merged sample of CRSP and Compustat between 1980 and 2019, and we restrict to observations for which data on sales, book equity, market capitalization, and stock returns are available. Column (1) reproduces the classic result that companies with high book-to-market ratios have higher future stock returns (the value premium). Column (2) is our main test of Equation (6.2), namely the return predictability is stronger for firms that have less persistent sales growth. The coefficient δ on the interaction term is indeed negative and significant. This finding is in line with our prediction: if the value premium is related to overreaction, such overreaction is more pronounced when firms' cash flows are more transitory. Columns (3) to (5) present robustness checks using various subsamples, including firms with market capitalization below and above the median and firms where we have at least 10 observations to estimate ρ_i. In all of these regressions, the t-statistic of the interaction term δ is very high, hovering between 5 and 6. Column (6) breaks down ρ_i into quintile dummies and shows that the effect of persistence is monotonic. The evidence in Table II is also consistent with earlier findings by Bouchaud et al. (2019), who show that companies with more persistent profits tend to exhibit stronger momentum (i.e., high past returns predict high future returns, which is often viewed as an indication of underreaction).

Table II: Value Premium and Persistence of Sales Growth

This table shows results of the following Fama-MacBeth regressions for firm i at fiscal year t:

$$r_{it+1} = \alpha + \beta BM_{it} + \gamma\rho_i + \delta BM_{it} \times \rho_i + \epsilon_{it},$$

where r_{it+1} is the annual stock return during fiscal year $t+1$, and BM_{it} is the book-to-market ratio of equity in the last day of fiscal year t. ρ_i ("persistence" in the table) is the autoregression coefficient of annual sales growth for firm i. The sample period 1980 to 2019. Each regression contains the persistence measure as a control (ρ_i in columns (1), (2), (3), (4), and (6); quintiles of ρ_i in column (5)). Column (1) shows the regression with the book-to-market ratio only. Columns (2) includes the interaction with the sales growth persistence ρ_i. Columns (3) to (5) perform robustness checks using various subsamples, including firms below and above the median market capitalization and firms where we have at least 10 observations to estimate ρ_i. Column (6) shows the monotonicity by splitting ρ_i into quintiles (with breakpoints defined every year). *** stands for 1% significance, and standard errors are reported in parentheses.

	Baseline (1)	All (2)	Small (3)	Large (4)	$N > 10$ (5)	Quintiles (6)
Book-to-Market (BM)	0.16***	0.18***	0.21***	0.14***	0.17***	0.24***
	(0.032)	(0.033)	(0.032)	(0.038)	(0.031)	(0.057)
BM × Persistence		-0.15***	-0.15***	-0.13***	-0.19***	
		(0.025)	(0.031)	(0.026)	(0.030)	
BM × Persistence Quintile 1						0.0019
						(0.048)
BM × Persistence Quintile 2						-0.054
						(0.048)
BM × Persistence Quintile 3						-0.069
						(0.050)
BM × Persistence Quintile 4						-0.10*
						(0.055)
BM × Persistence Quintile 5						-0.18***
						(0.057)
Observations	176,245	171,899	79,153	92,721	144,747	176,245
R^2	0.02	0.02	0.03	0.03	0.02	0.03

Taken together, results from stock returns presented above, along with results from analyst forecasts shown in Section 2.1, all point towards consistent patterns that align with our empirical facts.

7 Conclusion

Recent research using survey data from different sources points to varying degrees of biases in expectations. A key question is how to unify these different findings. To have a better understanding of how biases vary with the setting, we conduct a large-scale randomized experiment where participants forecast stable random processes. The experiment allows us to control the DGP and the relevant information sets. This is not feasible in survey data, which can give rise to major complications in interpreting results from survey data.

We find that forecasts display significant overreaction: they respond too much to recent observations. Furthermore, overreaction is particularly pronounced for less persistent processes and longer forecast horizons. We also find that standard specifications of commonly-used expectations models, estimated in our data, do not square with the variation in overreaction. Some predict too much overreaction when the process is transitory, while others predict too little.

We propose a framework to capture biases in expectations formation, where forecasters form estimates of the long-run mean of the process using a mix of the recent observation and past data. They balance these two sources of information depending on the setting, under the constraint that the utilization of past information is costly. As a result, forecasts adapt partially to the setting, but recent observations have a disproportionate influence, resulting in overreaction. Over-adjusting the estimates of the long-run mean in response to recent observations also implies that overreaction is more pronounced when the process is more transitory and the forecast horizon is longer. We estimate the model in our data and find that it closely matches how overreaction varies with process persistence. The model, when estimated on short-term forecasts, also predicts biases in long-term forecasts that closely match what we observe in the data.

Finally, our baseline model focuses on overreaction given the empirical evidence in our experiment. Nonetheless, the model can also be extended to allow for underreaction by introducing noisy signals, which could be a reason for underreaction observed in some survey data (Coibion and Gorodnichenko, 2012; Bouchaud et al., 2019). In this setting, the model maintains the prediction that the degree of overreaction should be relatively stronger when the process is less persistent. Taken together, we hope that the evidence and theory in this paper contribute to a systematic understanding of the findings on expectation biases.

References

Abarbanell, Jeffry and Victor Bernard, "Tests of Analysts' Overreaction/Underreaction to Earnings Information as an Explanation for Anomalous Stock Price Behavior," *Journal of Finance*, 1992.

Afrouzi, Hassan and Choongryul Yang, "Dynamic Rational Inattention and the Phillips Curve," Working Paper 2020.

Amromin, Gene and Steven A Sharpe, "From the Horse's Mouth: Economic Conditions and Investor Expectations of Risk and Return," *Management Science*, 2013, *60* (4), 845–866.

Andreassen, Paul B and Stephen J Kraus, "Judgmental Extrapolation and the Salience of Change," *Journal of Forecasting*, 1990, *9* (4), 347–372.

Angeletos, George-Marios, Zhen Huo, and Karthik A Sastry, "Imperfect Macroeconomic Expectations: Evidence and Theory," *NBER Macroeconomics Annual*, 2021, *35* (1), 1–86.

Asparouhova, Elena, Michael Hertzel, and Michael Lemmon, "Inference From Streaks in Random Outcomes: Experimental Evidence on Beliefs in Regime Shifting and the Law of Small Numbers," *Management Science*, 2009, *55* (11), 1766–1782.

Assenza, Tiziana, Te Bao, Cars Hommes, and Domenico Massaro, "Experiments on Expectations in Macroeconomics and Finance," *Research in Experimental Economics*, 2014, *17*, 11–70.

Baddeley, Alan, *Working Memory, Thought, and Action*, Vol. 45, Oxford University Press, 2007.

Baddeley, Alan D and Graham Hitch, "The Recency Effect: Implicit Learning with Explicit Retrieval?," *Memory & Cognition*, 1993, *21* (2), 146–155.

Barberis, Nicholas, Andrei Shleifer, and Robert Vishny, "A Model of Investor Sentiment," *Journal of Financial Economics*, 1998, *49* (3), 307–343.

_, Robin Greenwood, Lawrence Jin, and Andrei Shleifer, "X-CAPM: An Extrapolative Capital Asset Pricing Model," *Journal of Financial Economics*, 2015, *115*, 1–24.

Barrero, Jose Maria, "The micro and macro of managerial beliefs," *Journal of Financial Economics*, 2021.

Barrett, Lisa Feldman, Michele M Tugade, and Randall W Engle, "Individual Differences in Working Memory Capacity and Dual-Process Theories of the Mind," *Psychological Bulletin*, 2004, *130* (4), 553.

Ben-David, Itzhak, John R Graham, and Campbell R Harvey, "Managerial Miscalibration," *Quarterly Journal of Economics*, 2013, *128* (4), 1547–1584.

Benjamin, Daniel J, "Errors in Probabilistic Reasoning and Judgment Biases," *Handbook of Behavioral Economics: Applications and Foundations*, 2019, *2*, 69–186.

Beshears, John, James J Choi, Andreas Fuster, David Laibson, and Brigitte C Madrian, "What Goes Up Must Come Down? Experimental Evidence on Intuitive Forecasting," *American Economic Review*, 2013, *103* (3), 570–574.

Bloomfield, Robert and Jeffrey Hales, "Predicting the Next Step of a Random Walk: Experimental Evidence of Regime-Shifting Beliefs," *Journal of Financial Economics*, 2002, *65* (3), 397–414.

Bordalo, Pedro, John Conlon, Nicola Gennaioli, Spencer Kwon, and Andrei Shleifer, "Memory and Probability," Working Paper 2021.

_, Katherine Coffman, Nicola Gennaioli, Frederik Schwerter, and Andrei Shleifer, "Memory and Representativeness.," *Psychological Review*, 2020, *128* (1), 71.

_, Nicola Gennaioli, and Andrei Shleifer, "Diagnostic Expectations and Credit Cycles," *Journal of Finance*, 2018, *73* (1), 199–227.

_, _, and _, "Memory, Attention, and Choice," *Quarterly Journal of Economics*, 2020, *135* (3), 1399–1442.

_, _, Rafael La Porta, and Andrei Shleifer, "Diagnostic Expectations and Stock Returns," *Journal of Finance*, 2019, *74* (6), 2839–2874.

_, _, Yueran Ma, and Andrei Shleifer, "Overreaction in Macroeconomic Expectations," *American Economic Review*, 2020, *110* (9), 2748–82.

Bouchaud, Jean-Philippe, Philipp Krueger, Augustin Landier, and David Thesmar, "Sticky Expectations and the Profitability Anomaly," *Journal of Finance*, 2019, *74* (2), 639–674.

Cagan, Phillip, "The Monetary Dynamics of Hyperinflation," *Studies in the Quantity Theory of Money*, 1956.

Casler, Krista, Lydia Bickel, and Elizabeth Hackett, "Separate but Equal? A Comparison of Participants and Data Gathered via Amazon's MTurk, Social Media, and Face-To-Face Behavioral Testing," *Computers in Human Behavior*, 2013, *29* (6), 2156–2160.

Cavallo, Alberto, Guillermo Cruces, and Ricardo Perez-Truglia, "Inflation Expectations, Learning, and Supermarket Prices: Evidence From Survey Experiments," *American Economic Journal: Macroeconomics*, 2017, *9* (3), 1–35.

Coibion, Olivier and Yuriy Gorodnichenko, "What Can Survey Forecasts Tell Us About Information Rigidities?," *Journal of Political Economy*, 2012, *120*, 116–159.

_ and _, "Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts," *American Economic Review*, 2015, *105* (8), 2644–78.

Cover, Thomas M and Joy A Thomas, *Elements of Information Theory*, John Wiley & Sons, 1991.

Cowan, Nelson, *Attention and Memory: An Integrated Framework*, Oxford University Press, 1998.

— , "The Many Faces of Working Memory and Short-Term Storage," *Psychonomic Bulletin & Review*, 2017, *24* (4), 1158–1170.

— , "Working Memory: The Information You are Now Thinking of," in John H. Byrne, ed., *Learning and Memory: A Comprehensive Reference (Second Edition)*, second edition ed., Oxford: Academic Press, 2017, pp. 147–161.

da Silveira, Rava Azeredo and Michael Woodford, "Noisy Memory and Overreaction to News," in "AEA Papers and Proceedings," Vol. 109 2019, pp. 557–61.

— , Yeji Sung, and Michael Woodford, "Optimally Imprecise Memory and Biased Forecasts," Working Paper 2020.

D'Acunto, Francesco, "Identity, Overconfidence, and Investment Decisions," Working Paper 2015.

— and Michael Weber, "Memory and Beliefs: Evidence from the Field," Working Paper 2020.

d'Arienzo, Daniele, "Increasing Overreaction and Excess Volatility of Long Rates," Working Paper 2020.

De Bondt, Werner, "Betting on Trends: Intuitive Forecasts of Financial Risk and Return," *International Journal of Forecasting*, 1993, *9* (3), 355–371.

— and Richard H Thaler, "Do Security Analysts Overreact?," *American Economic Review*, 1990, pp. 52–57.

DellaVigna, Stefano and Devin Pope, "What Motivates Effort? Evidence and Expert Forecasts," *Review of Economic Studies*, 2017, *85* (2), 1029–1069.

— and — , "Predicting Experimental Results: Who Knows What?," *Journal of Political Economy*, 2018, *126* (6), 2410–2456.

Dwyer, Gerald P, Arlington W Williams, Raymond C Battalio, and Timothy I Mason, "Tests of Rational Expectations in a Stark Setting," *Economic Journal*, 1993, *103* (418), 586–601.

Enke, Benjamin, Frederik Schwerter, and Florian Zimmermann, "Associative Memory and Belief Formation," Working Paper 2021.

Evans, George and Seppo Honkapohja, *Learning and Expectations in Macroeconomics*, Princeton University Press, 2001.

Evans, Jonathan St BT, "Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition," *Annu. Rev. Psychol.*, 2008, *59*, 255–278.

Fan, Tony, Yucheng Liang, and Cameron Peng, "Belief Updating: Inference Versus Forecast Revision," Working Paper 2021.

Frydman, Cary and Gideon Nave, "Extrapolative Beliefs in Perceptual and Economic Decisions: Evidence of a Common Mechanism," *Management Science*, 2016, *63* (7), 2340–2352.

Fuster, Andreas, David Laibson, and Brock Mendel, "Natural Expectations and Macroeconomic Fluctuations," *Journal of Economic Perspectives*, 2010, *24* (4), 67–84.

Gabaix, Xavier, "Behavioral Inattention," *Handbook of Behavioral Economics*, 2018.

Gennaioli, Nicola, Yueran Ma, and Andrei Shleifer, "Expectations and Investment," *NBER Macroeconomics Annual*, 2016, *30* (1), 379–431.

Giglio, Stefano and Bryan Kelly, "Excess Volatility: Beyond Discount Rates," *Quarterly Journal of Economics*, 2018, *133* (1), 71–127.

Greenwood, Robin and Andrei Shleifer, "Expectations of Returns and Expected Returns," *Review of Financial Studies*, 2014, *27* (3), 714–746.

_ **and Samuel G Hanson**, "Waves in Ship Prices and Investment," *Quarterly Journal of Economics*, 2015, *130* (1), 55–109.

Hartzmark, Samuel M, Samuel Hirshman, and Alex Imas, "Ownership, Learning, and Beliefs," *Quarterly Journal of Economics*, 2021, *Forthcoming.*

Hey, John D, "Expectations Formation: Rational or Adaptive or ...?," *Journal of Economic Behavior & Organization*, 1994, *25* (3), 329–349.

Hirshleifer, David, Jun Li, and Jianfeng Yu, "Asset Pricing in Production Economies With Extrapolative Expectations," *Journal of Monetary Economics*, 2015, *76*, 87–106.

Ilut, Cosmin L and Rosen Valchev, "Economic Agents as Imperfect Problem Solvers," Working Paper 2020.

Kahana, Michael Jacob, *Foundations of Human Memory*, Oxford University Press, 2012.

Kahneman, Daniel and Amos Tversky, "Subjective Probability: A Judgment of Representativeness," *Cognitive Psychology*, 1972, *3* (3), 430–454.

Khaw, Mel Win, Ziang Li, and Michael Woodford, "Cognitive Imprecision and Small-Stakes Risk Aversion," *Review of Economic Studies*, 2018.

Kohlhas, Alexandre and Ansgar Walther, "Asymmetric Attention," *American Economic Review*, September 2021, *111* (9), 2879–2925.

Kuziemko, Ilyana, Michael I Norton, Emmanuel Saez, and Stefanie Stantcheva, "How Elastic Are Preferences for Redistribution? Evidence From Randomized Survey Experiments," *American Economic Review*, 2015, *105* (4), 1478–1508.

La Porta, Rafael, "Expectations and the Cross-Section of Stock Returns," *Journal of Finance*, 1996, *51* (5), 1715–1742.

Lakonishok, Josef, Andrei Shleifer, and Robert Vishny, "Contrarian Investment, Extrapolation and Risk," *Journal of Finance*, 1994, *49* (5), 1541–1578.

Lian, Chen, Yueran Ma, and Carmen Wang, "Low Interest Rates and Risk-Taking: Evidence From Individual Investment Decisions," *Review of Financial Studies*, 09 2018, *32* (6), 2107–2148.

Ma, Yueran, Tiziano Ropele, David Sraer, and David Thesmar, "A Quantitative Analysis of Distortions in Managerial Forecasts," Working Paper 2020.

Malmendier, Ulrike and Geoffrey Tate, "CEO Overconfidence and Corporate Investment," *Journal of Finance*, 2005, *60* (6), 2661–2700.

_ **and Stefan Nagel**, "Learning From Inflation Experiences," *Quarterly Journal of Economics*, 2016, *131* (1), 53–87.

Mankiw, Gregory and Ricardo Reis, "Sticky Information Versus Sticky Prices: A Proposal to Replace the New Keynesian Philips Curve," *Quarterly Journal of Economics*, 2002, *117* (4), 1295–1328.

Massey, Cade and George Wu, "Detecting Regime Shifts: The Causes of Under- and Overreaction," *Management Science*, 2005, *51* (6), 932–947.

Metzler, Lloyd A, "The Nature and Stability of Inventory Cycles," *Review of Economics and Statistics*, 1941, *23* (3), 113–129.

Nagel, Stefan and Zhengyang Xu, "Asset Pricing With Fading Memory," *Review of Financial Studies*, 2019, *Forthcoming.*

Neligh, Nathaniel, "Rational Memory with Decay," Working Paper 2020.

Nerlove, Marc, "Adaptive Expectations and Cobweb Phenomena," *Quarterly Journal of Economics*, 1958, *72* (2), 227–240.

Pope, Devin G and Maurice E Schweitzer, "Is Tiger Woods Loss Averse? Persistent Bias in the Face of Experience, Competition, and High Stakes," *American Economic Review*, 2011, *101* (1), 129–57.

Rabin, Matthew, "Inference by Believers in the Law of Small Numbers," *Quarterly Journal of Economics*, 2002, *117* (3), 775–816.

_ **and Dimitri Vayanos**, "The Gambler's and Hot-Hand Fallacies: Theory and Applications," *Review of Economic Studies*, 2010, *77* (2), 730–778.

Reimers, Stian and Nigel Harvey, "Sensitivity to Autocorrelation in Judgmental Time Series Forecasting," *International Journal of Forecasting*, 2011, *27* (4), 1196–1214.

Schmalensee, Richard, "An Experimental Study of Expectation Formation," *Econometrica*, 1976, pp. 17–41.

Sims, Christopher A, "Implications of Rational Inattention," *Journal of Monetary Economics*, 2003, *50* (3), 665–690.

Spillers, G, G Brewer, and N Unsworth, "Working Memory and Information Processing," *Encyclopedia of the Sciences of Learning, New York: Springer*, 2012, pp. 3474–3476.

Unsworth, Nash and Gregory J Spillers, "Working Memory Capacity: Attention Control, Secondary Memory, or Both? A Direct Test of the Dual-Component Model," *Journal of Memory and Language*, 2010, *62* (4), 392–406.

Wachter, Jessica A and Michael Jacob Kahana, "A Retrieved-Context Theory of Financial Decisions," Working Paper 2021.

Wang, Chen, "Under- and Over-Reaction in Yield Curve Expectations," Working Paper 2020.

Woodford, Michael, "Imperfect Common Knowledge and the Effects of Monetary Policy," *Knowledge, Information, and Expectations in Modern Macroeconomics*, 2003.

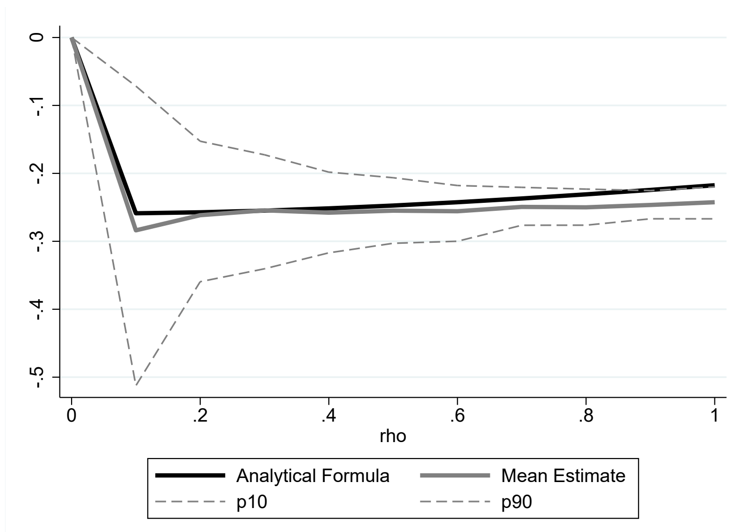
APPENDIX – FOR ONLINE PUBLICATION

A Appendix Figures and Tables

Figure A.1: Estimation Error: Error-Revision Coefficient and Implied Persistence

This figure shows simulation results on the error-revision coefficient and the implied persistence. We start by simulating 10 datasets of 45 participants each, where each participant makes 40 forecasts of an AR(1) process. Each of the 10 dataset has one level of the AR(1) persistence ρ, which goes from 0 to 1. In each dataset, participants make forecasts using the diagnostic expectations model: $F_t x_{t+h} = \rho^h x_t + 0.4\rho^h \epsilon_t$, where x_t is the process realization and ϵ_t is the innovation. In Panel A, for each level of ρ, we estimate the error-revision coefficient b from the following regression: $x_{t+1} - F_t x_{t+1} = c + b(F_t x_{t+1} - F_{t-1} x_{t+1}) + u_{t+1}$. The dark solid line shows the theoretical prediction (Bordalo et al., 2020c). The light solid line shows the average coefficient from 200 simulations. The dashed lines show the 90% confidence bands from the simulations. In Panel B, we implement the same procedure and report the implied persistence coefficient $\hat{\rho}$ estimated from the regression: $F_t x_{t+1} = cst + \hat{\rho} x_t + v_{t+1}$. The dark solid line shows the theoretical prediction based on diagnostic expectations. The light solid line shows the average coefficient from 200 simulations. The dashed lines show the 90% confidence bands from the simulations. The standard errors are very tight so the three lines lie on top of one another.

Panel A. Error-Revision Coefficient



Panel B. Implied Persistence

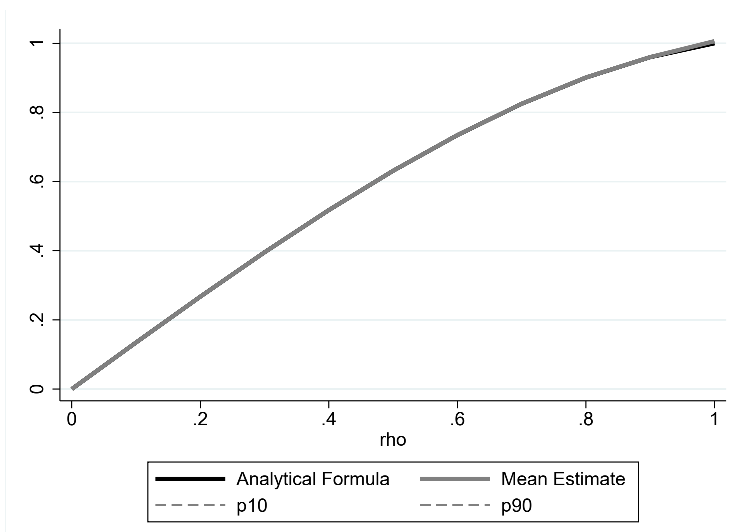


Figure A.2: Prediction Screen

This figure shows a screenshot of the prediction task. The green dots indicate past realizations of the statistical process. In each round t, participants are asked to make predictions about two future realizations $F_t x_{t+1}$ and $F_t x_{t+2}$. They can drag the mouse to indicate $F_t x_{t+1}$ in the purple bar and indicate $F_t x_{t+2}$ in the red bar. Their predictions are shown as yellow dots. The grey dot is the prediction of x_{t+1} from the previous round ($F_{t-1} x_{t+1}$); participants can see it but cannot change it. After they have made their predictions, participants click "Make Predictions" and move on to the next round. The total score is displayed on the top left corner, and the score associated with each of the past prediction (if the actual is realized) is displayed at the bottom.

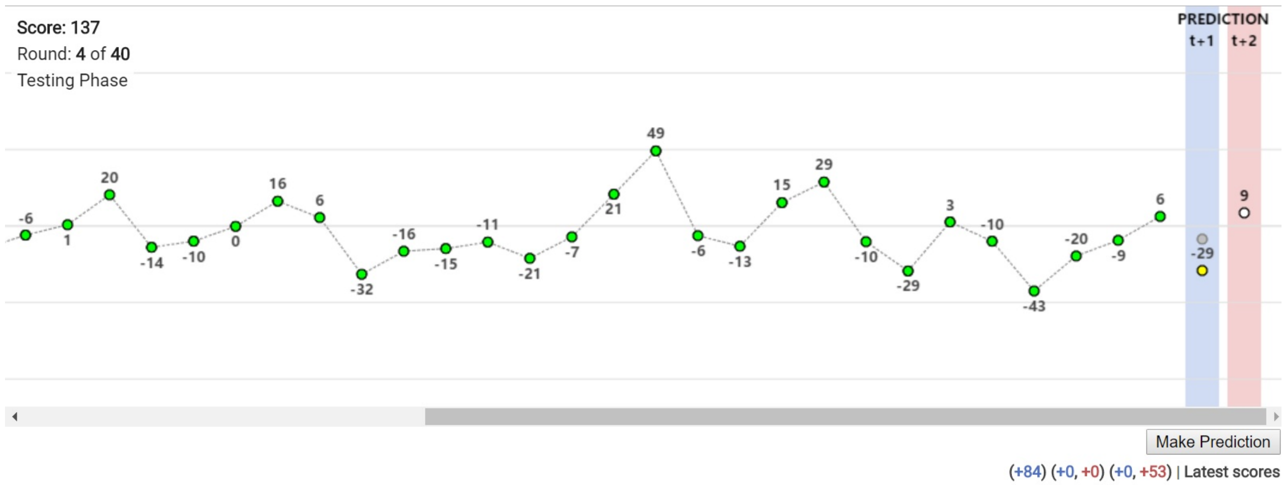


Figure A.3: Implied Persistence Relative to Actual Persistence

We compute the implied persistence ρ_1^s from $F_{it}x_{t+1} = c + \rho_{s,1}x_t + u_{it}$ for each level of AR(1) persistence ρ. The y-axis plots the implied persistence relative to the actual persistence $\zeta(\rho, 1) = \rho_{s,1}/\rho$, i.e., the measure of overreaction, and the x-axis plots the AR(1) persistence ρ. The line at one is the FIRE benchmark.

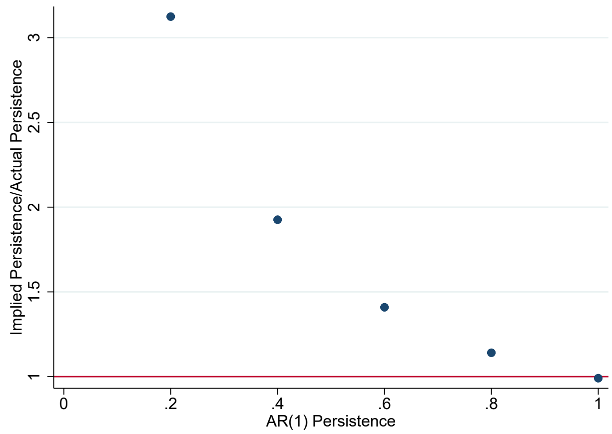
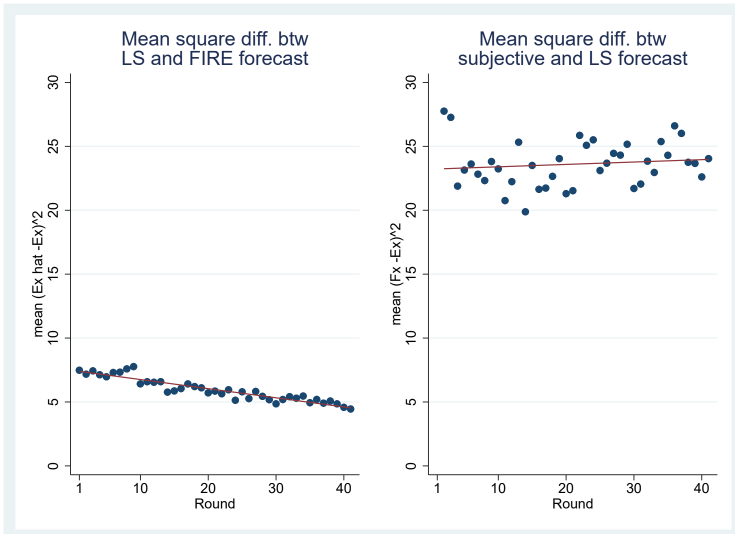


Figure A.4: Distance between Subjective Forecasts and Rational Expectations

The top left panel shows the root mean squared difference between in-sample least square expectations and full information rational expectations (FIRE). The top right panel shows the root mean squared difference between participants' actual subjective forecasts and the least square forecasts. The data use all conditions in Experiment 1. The bottom panel shows the implied persistence of least square forecasts for each level of ρ, which is the regression coefficient of the least square forecast on x_t.

Panel A. Least Square Forecasts vs. FIRE and Subjective Forecasts



Panel B. Implied Persistence of Least Square Forecasts

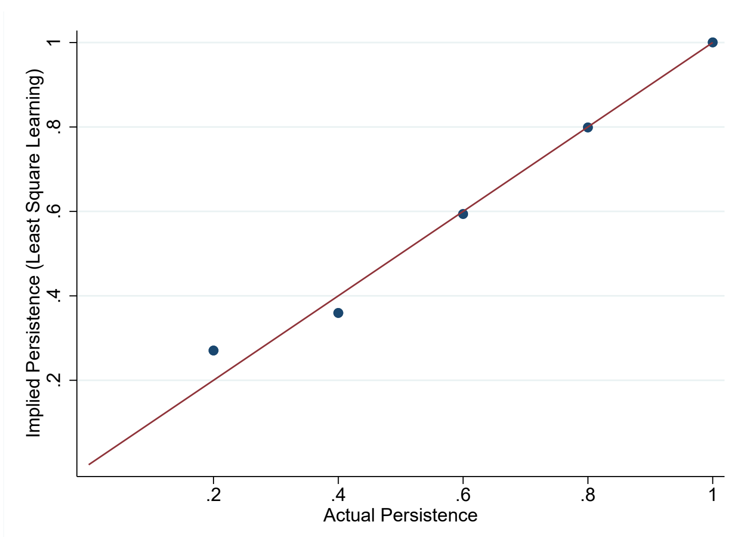
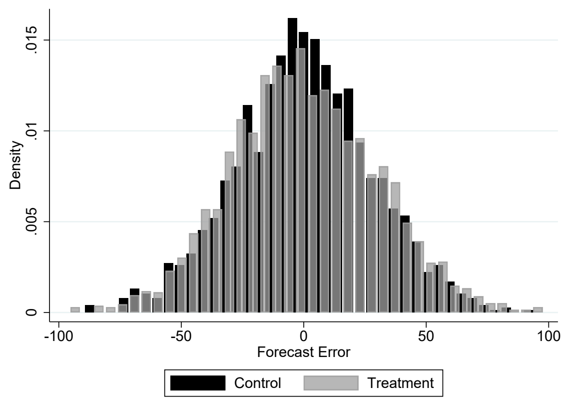


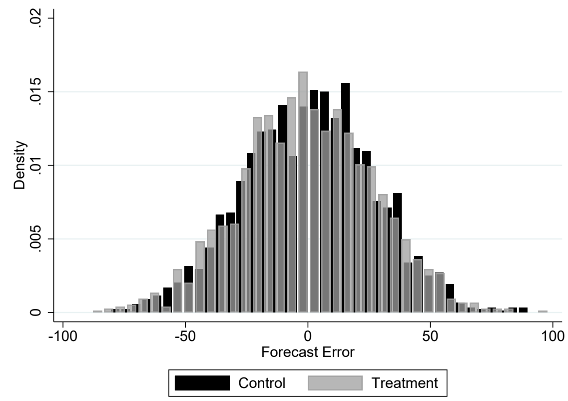
Figure A.5: Knowledge of Linear DGP and the Distribution of Forecasts

We use the data from Experiment 3 (MIT EECS), with 204 MIT undergraduates randomly assigned to AR(1) processes with $\rho = 0.2$ or $\rho = 0.6$. 94 randomly selected participants were told that the process is a stable random process (control group), while 110 were told that the process is an AR(1) with fixed μ and ρ (treatment group). Panel A shows the distributions of the forecast error $x_{t+1} - F_t x_{t+1}$ for both treated and control groups. Panel B shows binscatter plots of the forecast error as a function of the latest realization x_t.

Panel A. Distribution of Forecast Error ($x_{t+1} - F_t x_{t+1}$)

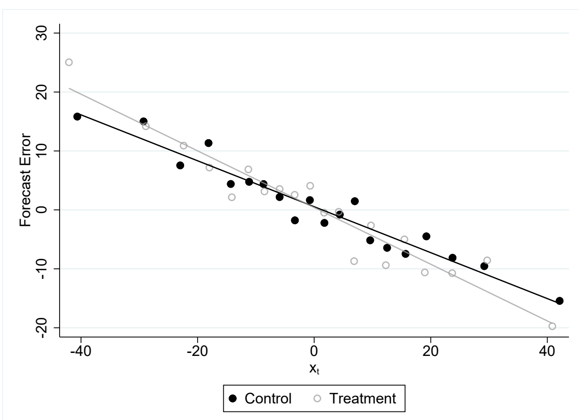


$\rho = 0.2$

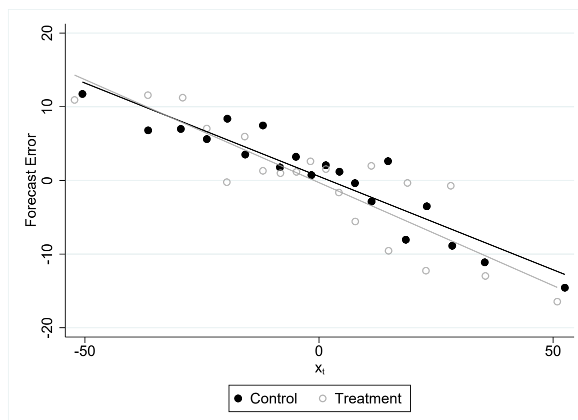


$\rho = 0.6$

Panel B. Forecast Error Conditional on x_t



$\rho = 0.2$



$\rho = 0.6$

Figure A.6: Comparison of Experiment 3 (MIT EECS) and Experiment 1 (Baseline)

This figure shows the implied persistence (i.e., regression coefficient of the forecast $F_t x_{t+1}$ on x_t) in Experiment 3 (MIT EECS) and Experiment 1 (MTurk baseline).

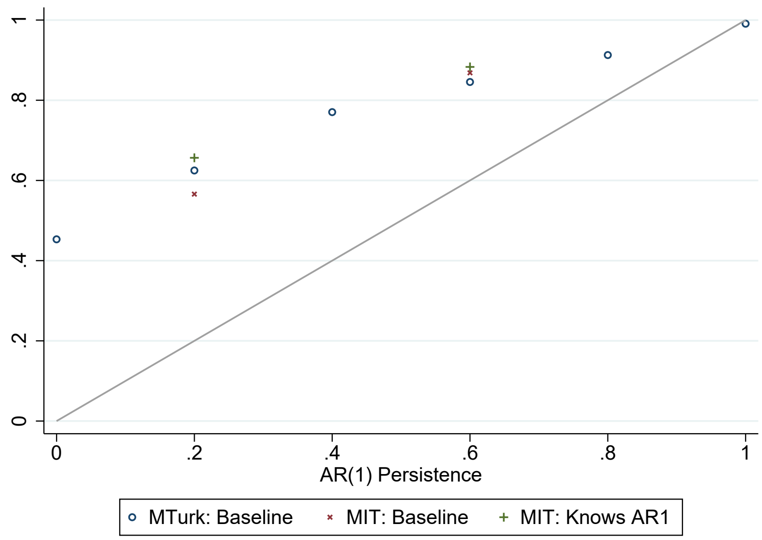
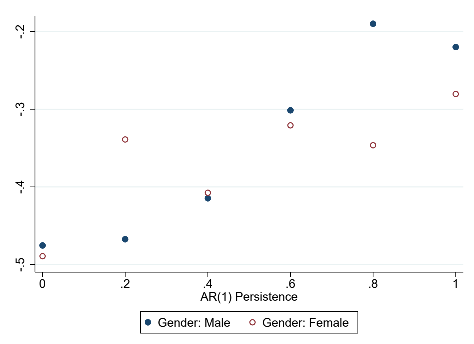


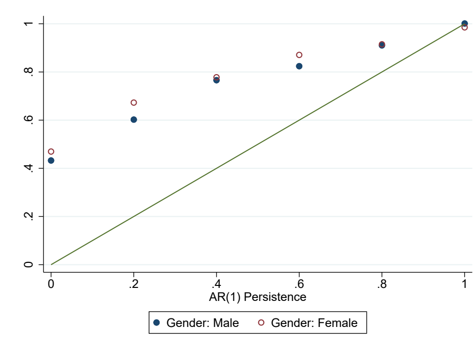
Figure A.7: Overreaction and Persistence of Process: Results by Demographics

This figure plots the error-revision coefficient and the implied persistence for each level of AR(1) persistence, estimated in different demographic groups. In Panel A, the solid dots represent results for male participants and the hollow dots represent results for female participants. In Panel B, the solid dots represent results for participants younger than 35 and the hollow dots represent results for participants older than 35. In Panel C, the solid dots represent results for participants with high school degrees, and the hollow dots represent results for participants with college and above degrees.

Panel A. By Gender: Male vs. Female

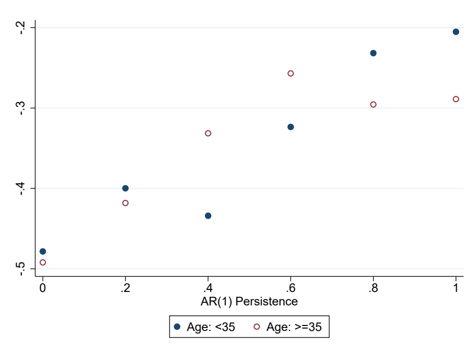


Error-Revision Coefficient by Gender

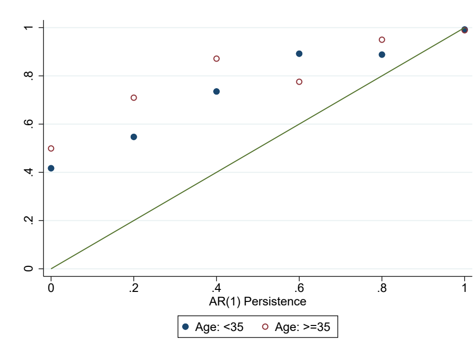


Implied Persistence by Gender

Panel B. By Age: Below 35 vs. Above 35

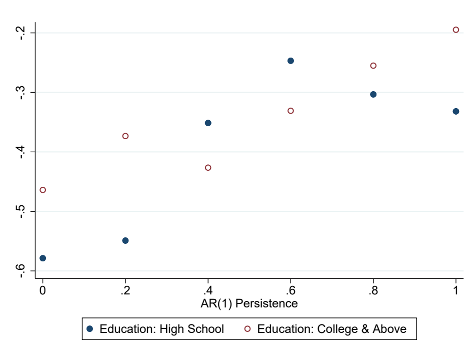


Error-Revision Coefficient by Age

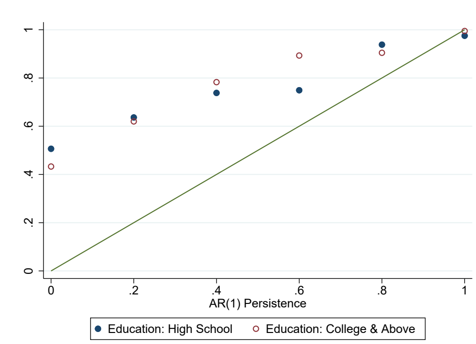


Implied Persistence by Age

Panel C. By Education: High School vs. College and Above



Error-Revision Coefficient by Education



Implied Persistence by Education

Figure A.8: Error-Revision Coefficient: Data vs Models

For each level of ρ, we regress the model-based forecast error $x_{t+1} - \widehat{F_t^m x_{t+1}}$ on the model-based forecast revision $\widehat{F_t^m x_{t+1}} - \widehat{F_{t-1}^m x_{t+1}}$. The dots report the regression coefficient obtained for each model m and each level of ρ. The solid line reports the error-revision coefficient in the experimental data, as in Figure II, Panel A.

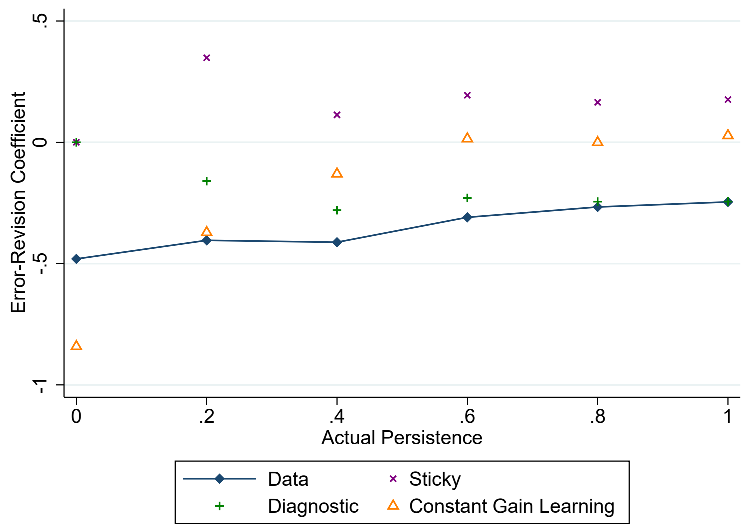
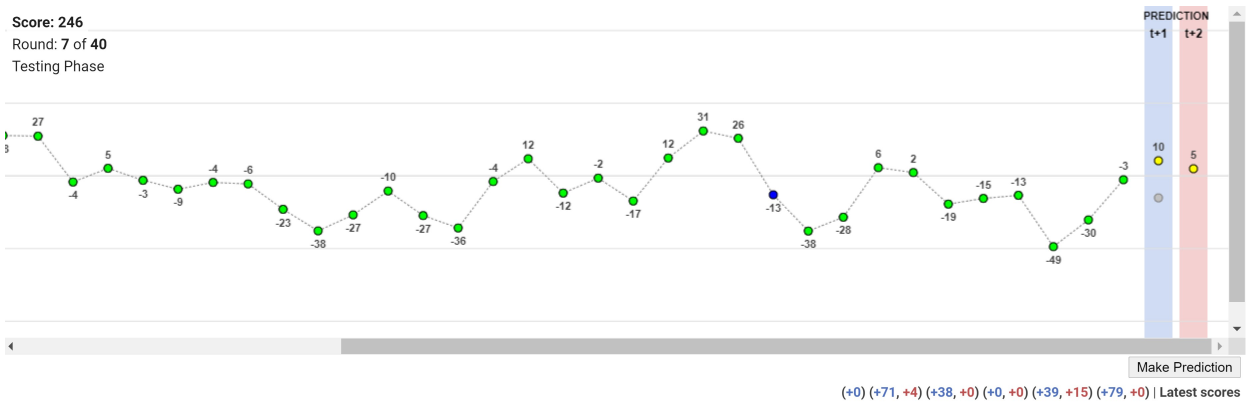


Figure A.9: Prediction Screen for Additional Experiments

This figure shows the screenshot of the prediction task for addittional treatment conditions in Experiment 4. Panel A shows the condition where we require participants to click on x_{t-10} (the dot in blue) before making the prediction. Panel B shows the condition where we include a red line at zero.

Panel A. Click x_{t-10}



Panel B. Show Red Line at 0

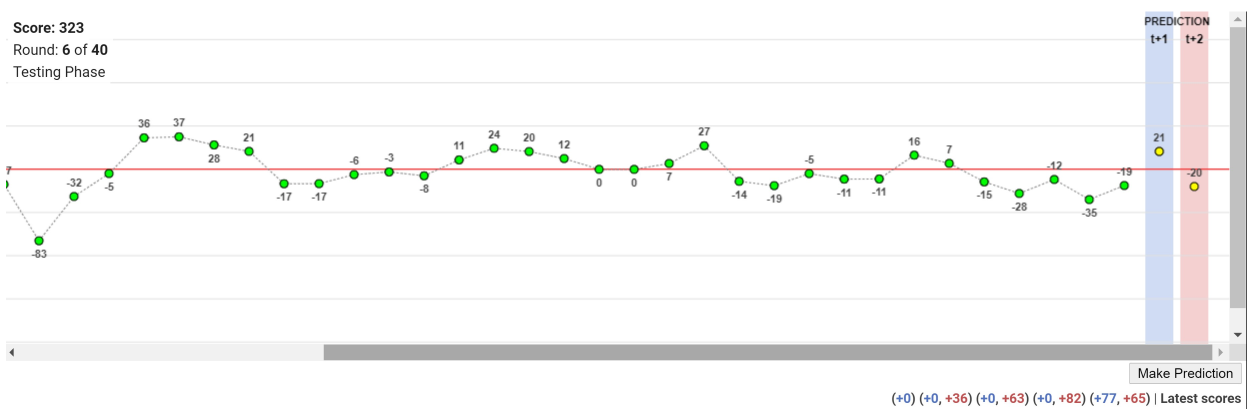
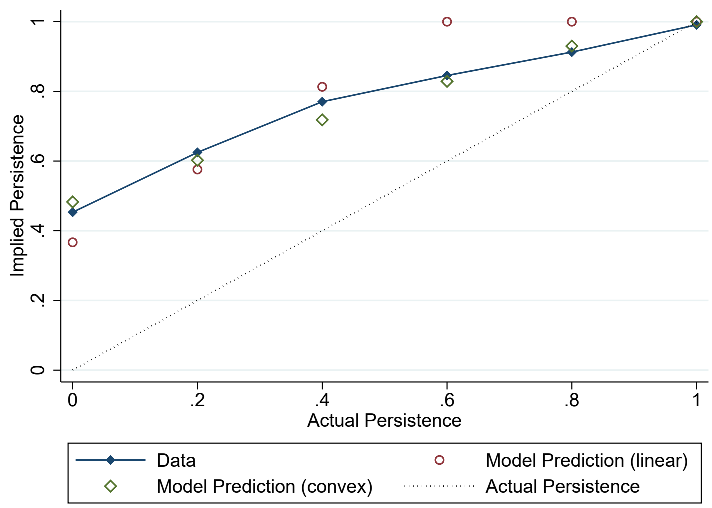


Figure A.10: Model Functional Form: Robustness Checks

This figure shows the model fit under alternative model specifications of the cost function, for $h = 1$ in Panel A and $h = 5$ in Panel B. The red dots represent the implied persistence from our model when $\gamma = 1$, and the green diamonds represent result from our model when we do a full grid search for γ. The blue line represents the value observed in the forecast data.

Panel A. $h = 1$



Panel B. $h = 5$

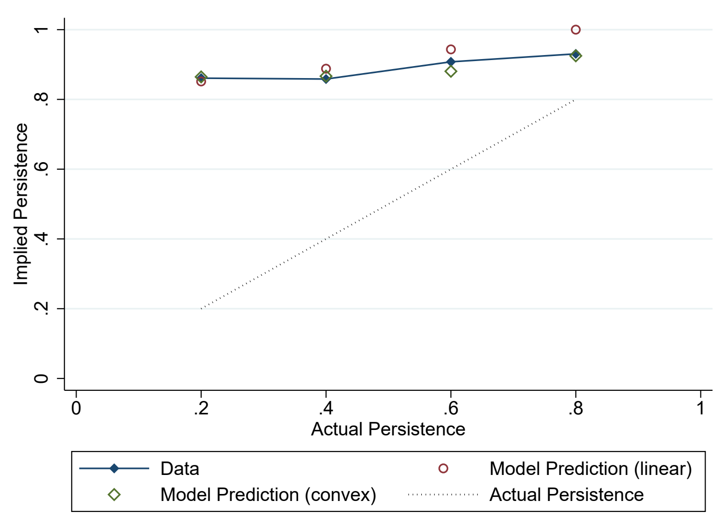
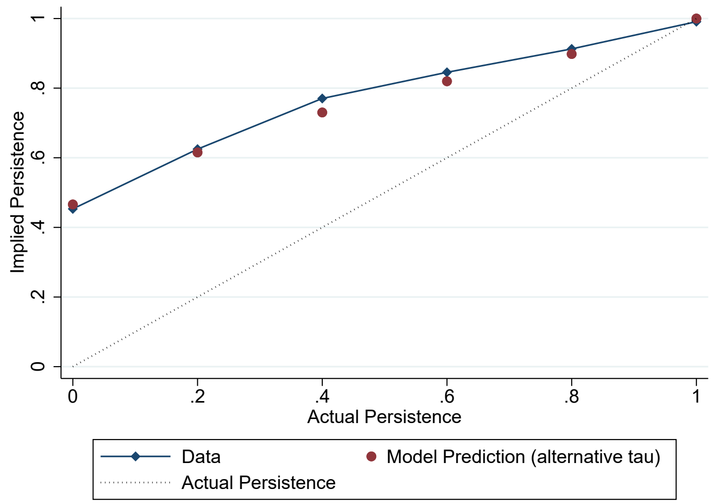


Figure A.11: Model Functional Form: Robustness Checks

This figure shows the model fit under the alternative formulation of $\underline{\tau}$, as discussed in Section 6.1, for $h = 1$ in Panel A and $h = 5$ in Panel B. The red dots represent the implied persistence from our model, and the blue line represents the value observed in the forecast data.

Panel A. $h = 1$



Panel B. $h = 5$

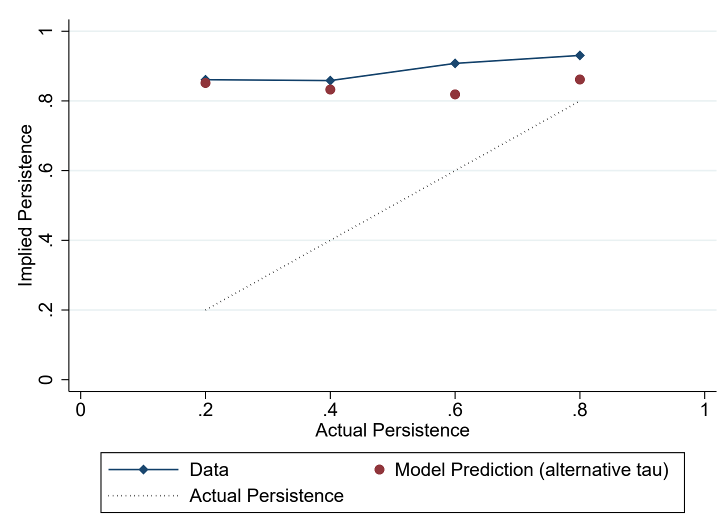


Table A.1: Experimental Literature on Expectations Formation

This table summarizes the experimental literature on forecasts of stochastic processes. The first column lists the authors and the date of publication. Column (2) displays the number of participants. Column (3) shows the number of process realizations shown at the beginning of the experiment. Column (4) reports the number of rounds of forecasts each participant has to make. Column (5) describes the process. Most of the time, it is an AR(1). In one case, it is an exponentially growing process. In another case, it is an integrated moving average. Column (6) shows that nearly all experiments feature some form of monetary incentives. Column (7) shows the forecast horizon requested. The last column describes the models tested.

(1) Paper	(2) # of Participants	(3) # of History	(4) # of Predictions	(5) Process	(6) Monetary Incentives	(7) Forecast Horizon	(8) Model Tested
Schmalensee (1976)	23	25	28	$\rho \approx 1$	Yes	1-5	Adaptive +Extrap.
Andreassen and Kraus (1990)	77	5	5	$e^{\alpha t}$	No	1	Extrap.
De Bondt (1993)	27	48	2	$\rho \approx 1$	Weak	7,13	Extrap.
Dwyer et al. (1993)	70	30	40	$\rho = 1$	Yes	1	Adaptive
Hey (1994)	48	50	40	$\rho \in \{0.1, 0.5, 0.8, 0.9\}$	Yes	1	Adaptive
Bloomfield and Hales (2002)	38	9	1	$\rho \approx 1$	Yes	1	BSV
Asparouhova, Hertzel and Lemmon (2009)	92	100	100	$\rho \approx 1$	Yes	1	BSV vs Rabin
Reimers and Harvey (2011)	2,434	50	Varies	$\rho \in \{0, 0.4, 0.8\}$	Yes	1	N/A
Beshears et al. (2013)	98	100k	60	ARIMA(0,1,10), ARIMA(0,1,50)	Yes	1	Natural Expec.
Frydman and Nave (2016)	38	10	400	$\rho \approx 1$	Yes	1	Extrap.
This Paper	1,600+	40	40	$\rho \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$	Yes	1,2,5, 10	Multiple

Table A.2: Summary of Conditions

This table provides a summary of the experiments we conducted. Each panel describes one experiment, and each line within a panel corresponds to one treatment condition. Columns (1) to (3) show the parameters of the AR(1) process $x_{t+1} = \mu + \rho x_t + \epsilon_{t+1}$. Participants are only allowed to participate once.

	Condition	(1) Persistence ρ	(2) Mean μ	(3) Conditional Vol σ_ϵ	(4) # of Participants
Panel A: Experiment 1 – Baseline, MTurk					
A1	Baseline	0	0	20	32
A2	Baseline	0.2	0	20	32
A3	Baseline	0.4	0	20	36
A4	Baseline	0.6	0	20	39
A5	Baseline	0.8	0	20	28
A6	Baseline	1	0	20	40
Panel B: Experiment 2 – Long Horizon, MTurk					
B1	Horizon: F1 + F5	0.2	0	20	41
B2	Horizon: F1 + F5	0.4	0	20	26
B3	Horizon: F1 + F5	0.6	0	20	31
B4	Horizon: F1 + F5	0.8	0	20	30
Panel C: Experiment 3 – DGP Information, MIT EECS					
C1	Baseline	0.2	0	20	42
C2	Baseline	0.6	0	20	52
C3	Display DGP is AR(1)	0.2	0	20	70
C4	Display DGP is AR(1)	0.6	0	20	40
Panel D: Experiment 4 – Additional Test, MTurk					
D11	Baseline	0	0	20	41
D12	Baseline	0.2	0	20	36
D13	Baseline	0.4	0	20	34
D14	Baseline	0.6	0	20	26
D15	Baseline	0.8	0	20	28
D16	Baseline	1	0	20	26
D21	Red Line at 0	0	0	20	34
D22	Red Line at 0	0.2	0	20	32
D23	Red Line at 0	0.4	0	20	24
D24	Red Line at 0	0.6	0	20	36
D25	Red Line at 0	0.8	0	20	39
D26	Red Line at 0	1	0	20	33
D31	Click x_{t-10}	0	0	20	23
D32	Click x_{t-10}	0.2	0	20	30
D33	Click x_{t-10}	0.4	0	20	28
D34	Click x_{t-10}	0.6	0	20	25
D35	Click x_{t-10}	0.8	0	20	28
D36	Click x_{t-10}	1	0	20	27
D41	Horizon: F1 + F10	0	0	20	27
D42	Horizon: F1 + F10	0.2	0	20	27
D43	Horizon: F1 + F10	0.4	0	20	30
D44	Horizon: F1 + F10	0.6	0	20	26
D45	Horizon: F1 + F10	0.8	0	20	36
D46	Horizon: F1 + F10	1	0	20	38

Table A.3: Summary Statistics

Panel A describes demographics of participants. Panel B reports basic experimental statistics, including the total score, the total bonus (incentive payments) paid in US dollars, the overall time taken to complete the experiment, and the time taken to complete the forecasting part (the main part).

Panel A. Participant Demographics

	Experiment 1		Experiment 2		Experiment 3		Experiment 4	
	Obs.	%	Obs.	%	Obs.	%	Obs.	%
Gender: Female	90	43.5	61	47.7	116	56.9	316	43.1
Gender: Male	117	56.5	67	52.3	88	43.1	418	56.9
Age: <= 25	30	14.5	18	14.1	197	96.6	62	8.4
Age: 25-45	138	66.7	89	69.5	7	3.4	500	68.1
Age: 45-65	35	16.9	20	15.6	0	0.0	156	21.3
Age: 65+	4	1.9	1	0.8	0	0.0	16	2.2
Education: Grad School	20	9.7	18	14.1	0	0.0	170	23.2
Education: College	132	63.8	74	57.8	204	100.0	426	58.0
Education: High School	55	26.6	36	28.1	0	0.0	133	18.1
Education: Below/Other	0	0.0	0	0.0	0	0.0	5	0.7
Invest. Exper.: Extensive	7	3.4	3	2.3	2	1.0	77	10.5
Invest. Exper.: Some	58	28.0	29	22.7	21	10.3	258	35.1
Invest. Exper.: Limited	71	34.3	56	43.8	43	21.1	232	31.6
Invest. Exper.: None	71	34.3	40	31.3	138	67.6	167	22.8
Taken Stat Class: No	117	56.5	80	62.5	0	0.0	361	49.2
Taken Stat Class: Yes	90	43.5	48	37.5	204	100.0	373	50.8

Panel B. Experimental Statistics

	Mean	p25	p50	p75	SD	*N*
Experiment 1						
Total Forecast Score	2,004	1,690	1,990	2,335	461.93	207
Bonus ($)	3.34	2.82	3.32	3.89	0.77	207
Total Time (min)	18.01	10.92	13.11	21.85	11.34	207
Forecast Time (min)	6.80	4.54	5.66	7.79	3.53	207
Experiment 2						
Total Forecast Score	1,843	1,588	1,820	2,138	463.38	128
Bonus ($)	3.07	2.65	3.04	3.56	0.77	128
Total Time (min)	15.82	8.74	13.11	19.66	9.80	128
Forecast Time (min)	6.70	4.54	6.02	7.58	3.17	128
Experiment 3						
Total Forecast Score	2,071	1,755	2,046	2,326	429.59	204
Bonus ($)	8.63	7.31	8.53	9.69	1.79	204
Total Time (min)	18.45	6.55	10.92	13.11	37.67	204
Forecast Time (min)	8.78	4.03	5.09	7.46	19.72	204
Experiment 4						
Total Forecast Score	1,767	1,422	1,812	2,174	610.23	734
Bonus ($)	2.95	2.37	3.02	3.62	1.02	734
Total Time (min)	15.75	8.74	13.11	19.66	10.00	734
Forecast Time (min)	7.88	4.79	6.50	9.22	4.97	734

Table A.4: Effect of Knowing the Process is AR(1)

This table reports the implied persistence in Experiment 3 among MIT EECS students. Participants are randomly assigned to $\rho = 0.2$ and $\rho = 0.6$. In addition, half of them are randomly assigned to the baseline control condition (control) where the process is described as a stable random process, while the other half are assigned to the treatment condition where they are told that the process is a fixed and stationary AR(1) process.

	Baseline Condition	Knows AR(1)	Difference (p-value)
$\rho = .2$	0.56	0.65	0.14
$\rho = .6$	0.86	0.88	0.71

Table A.5: Estimation of Expectation Models

This table reports estimation of eight expectation formation models. Each model is described by an equation and a parameter, highlighted in bold. Estimations are based on pooled data from all conditions of Experiment 1 (i.e., with $\rho \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$). All models except constant gain learning and FIRE (which has no parameter) are estimated using constrained least squares. We cluster standard errors at the individual level. The imperfect memory model is estimated by minimizing, over the decay parameter, the mean squared deviation between predicted and realized forecasts. We then estimate standard errors for this model by block-bootstrapping forecasters. The parameter estimate is reported in the third column, along with standard errors in the fourth column. In the fifth column, we report the mean squared error of each model, as a fraction of the sample variance of forecast. Since forecasts in the $\rho = 1$ condition are mechanically much more variable than the forecasts in the $\rho = 0$ condition, we report here the average of this ratio across conditions. This avoids giving too much weight to the low variance (low ρ) conditions.

Model	Equation	Parameter Estimate	Standard Error	Mean MSE / $\text{var}F_t x_{t+1}$
Panel A : Backward-Looking Models				
Adaptive	$F_t x_{t+1} = \boldsymbol{\delta} F_{t-1} x_t + (1 - \boldsymbol{\delta}) x_t$	0.17***	(0.04)	0.53
Extrapolative	$F_t x_{t+1} = (1 + \boldsymbol{\phi}) x_t - \boldsymbol{\phi} x_{t-1}$	-0.07***	(0.02)	0.56
Panel B : Forward-Looking Models				
FIRE	$F_t x_{t+1} = E_t x_{t+1}$	-	-	0.58
Sticky/noisy information	$F_t x_{t+1} = \boldsymbol{\lambda} F_{t-1} x_{t+1} + (1 - \boldsymbol{\lambda}) E_t x_{t+1}$	0.14***	(0.04)	0.56
Diagnostic	$F_t x_{t+1} = E_t x_{t+1} + \boldsymbol{\theta}(E_t x_{t+1} - E_{t-1} x_{t+1})$	0.34***	(0.04)	0.57
Constant gain learning	Rolling regression at t w/ weights: $w_s^t = \frac{1}{\boldsymbol{\kappa}^{t-s}}$	1.06***	(0.01)	0.56

Table A.6: Bias and Lagged Realizations

This table shows regressions of the forecast $F_t x_{t+1}$ (Panel A) or deviation from the rational benchmark $\rho x_t - F_t x_{t+1}$ (Panel B) on lags of realizations x_{t-k}. The data comes from Experiment 1.

Panel A: LHS is $F_t x_{t+1}$

$\rho =$	0	0.2	0.4	0.6	0.8	1
	(1)	(2)	(3)	(4)	(5)	(6)
x_t	0.44***	0.61***	0.78***	0.87***	1.01***	1.12***
	(0.05)	(0.06)	(0.04)	(0.05)	(0.05)	(0.05)
x_{t-1}	0.04	-0.04	-0.08**	-0.08	-0.13**	-0.15***
	(0.04)	(0.04)	(0.04)	(0.05)	(0.06)	(0.05)
x_{t-2}	-0.00	0.04	-0.01	0.02	-0.01	-0.05
	(0.04)	(0.03)	(0.04)	(0.05)	(0.04)	(0.04)
x_{t-3}	0.09**	0.04	0.08***	-0.01	0.03	0.04
	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)
x_{t-4}	0.08**	0.02	-0.04	0.07**	-0.01	0.03
	(0.03)	(0.04)	(0.03)	(0.03)	(0.02)	(0.03)
Observations	1,312	1,312	1,476	1,599	1,148	1,640
R^2	0.15	0.28	0.35	0.46	0.73	0.98

Panel B: LHS is $\rho x_t - F_t x_{t+1}$

$\rho =$	0	0.2	0.4	0.6	0.8	1
	(1)	(2)	(3)	(4)	(5)	(6)
x_t	-0.44***	-0.41***	-0.38***	-0.27***	-0.21***	-0.12**
	(0.05)	(0.06)	(0.04)	(0.05)	(0.05)	(0.05)
x_{t-1}	-0.04	0.04	0.08**	0.08	0.13**	0.15***
	(0.04)	(0.04)	(0.04)	(0.05)	(0.06)	(0.05)
x_{t-2}	0.00	-0.04	0.01	-0.02	0.01	0.05
	(0.04)	(0.03)	(0.04)	(0.05)	(0.04)	(0.04)
x_{t-3}	-0.09**	-0.04	-0.08***	0.01	-0.03	-0.04
	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)
x_{t-4}	-0.08**	-0.02	0.04	-0.07**	0.01	-0.03
	(0.03)	(0.04)	(0.03)	(0.03)	(0.02)	(0.03)
Observations	1,312	1,312	1,476	1,599	1,148	1,640
R^2	0.15	0.15	0.11	0.07	0.06	0.02

Table A.7: Model Fit

This table shows the MSE between ρ_h^s in the model in columns (1), (3), and (5), and the MSE between $F_t x_{t+h}$ implied by the model and $F_t x_{t+h}$ in the data in columns (2), (4), (6). Columns (1) and (2) report results for the 1-period forecast; columns (3) and (4) report results for the 2-period forecast; columns (5) and (6) report results for the 5-period forecast. The adaptive expectations model is: $F_t x_{t+1} = \delta x_t + (1-\delta) F_{t-1} x_t$. The traditional extrapolative expectations model is: $F_t x_{t+1} = x_t + \phi(x_t - x_{t-1})$. The sticky expectations model is: $F_t x_{t+h} = (1-\lambda)\rho^h x_t + \lambda F_{t-1} x_{t+h} + \epsilon_{it,h}$. The diagnostic expectations model is: $F_t x_{t+h} = E_t x_{t+h} + \theta(E_t x_{t+h} - E_{t-1} x_{t+h})$. The constant gain learning model is: $F_t x_{t+h} = \widehat{E}_t x_{t+h} = a_{t,h} + \sum_{k=0}^{k=n} b_{k,h,t} x_{t-k}$.

Forecast Horizon	$h=1$		$h=2$		$h=5$		$h=10$	
MSE Type	ρ_h^s (1)	Forecast (2)	ρ_h^s (3)	Forecast (4)	ρ_h^s (5)	Forecast (6)	ρ_h^s (7)	Forecast (8)
Current Model	0.003	496.1	0.001	719.2	0.001	691.0	0.001	2176.4
Adaptive	0.035	495.7	.	.	.	.	.	.
Extrapolative	0.064	527.3	.	.	.	.	.	.
Sticky	0.117	556.2	0.140	786.1	0.197	814.6	0.310	2304.9
Diagnostic	0.069	521.2	0.115	758.0	0.177	803.3	0.302	2338.2
Constant Gain Learning	0.067	526.8	0.039	749.5	0.033	736.3	0.022	2454.9

B Proofs

B.1 Standard Errors of Error-Revision Coefficient

Proposition B.1. Assume a univariate regression of centered variables:

$$y_i = \beta x_i + u_i.$$

Then, the standard error of the OLS estimate of β is given by:

$$\boxed{s.d.\left(\widehat{\beta} - \beta\right) \approx \frac{1}{\sqrt{N}} \left(\frac{\text{var} y_i}{\text{var} x_i} - \beta^2 \right)^{1/2}}.$$

Proof. The OLS estimator of β is given by:

$$\widehat{\beta} = \frac{\frac{1}{N} \sum_i x_i y_i}{\frac{1}{N} \sum_i x_i^2} = \beta + \frac{\frac{1}{N} \sum_i x_i u_i}{\frac{1}{N} \sum_i x_i^2}.$$

Hence,

$$\sqrt{N}(\widehat{\beta} - \beta) = \frac{\sqrt{N} \frac{1}{N} \sum_i x_i u_i}{\frac{1}{N} \sum_i x_i^2}.$$

By virtue of the central limit theorem, we have:

$$\sqrt{N} \frac{1}{N} \sum_i x_i u_i \to N(0, \text{var}(x_i u_i)),$$

while

$$\frac{1}{N} \sum_i x_i^2 \to \text{var} x_i.$$

This ensures that:

$$\sqrt{N}(\widehat{\beta} - \beta) \to N(0, \underbrace{\frac{\text{var}(x_i u_i))}{(\text{var}(x_i))^2}}_{= \frac{\text{var} u_i}{\text{var} x_i}}).$$

Note that the asymptotic variance can be rewritten as:

$$\begin{aligned} \frac{\text{var} u_i}{\text{var} x_i} &= \frac{\text{var} y_i + \beta^2 \text{var} x_i - 2\beta \text{cov}(x_i, y_i)}{\text{var} x_i} \\ &= \frac{\text{var} y_i}{\text{var} x_i} - \beta^2. \end{aligned}$$

□

Evidently, this ratio is bigger when the variance of x_i is smaller.
For the error-revision coefficient, it can easily be shown that:

$$\frac{\text{var} y_i}{\text{var} x_i} = \frac{(1 + \rho^2 \theta^2)}{\rho^2 \left((1 + \theta)^2 + \theta^2 \rho^2 \right)} \to +\infty \text{ as } \rho \to 0$$

This makes it clear that the error-revision coefficient does not work well for small ρ because the right-hand-side variable has a small variance, which makes it hard to estimate λ precisely.

On the other hand, measuring overreaction using the implied persistence does not have this problem as the variance of the right-hand-side variable is just the variance of the process itself, which is non-zero.

B.2 Lemma 1

Proof. The agent has two decisions. First, she decides what information to utilize (chooses $S_t \subseteq \mathcal{A}_t$). Second, she chooses the optimal forecast $F_t x_{t+h}$ given the σ-algebra induced by S_t. We solve this backwards. Specifically, we characterize the optimal forecast for any choice of S_t and then solve for the optimal S_t given the optimal forecast that it implies.

It is straightforward to see that with a quadratic loss function the optimal forecast for a given choice of S_t is simply the unbiased expectation of x_{t+h} conditional on S_t. Formally, let $F_t^* x_{t+h}(S_t)$ denote the optimal forecast of the agent under S_t, then

$$F_t^* x_{t+h}(S_t) \equiv \arg\min_{F_t x_{t+h}} \mathbb{E}[(F_t x_{t+h} - x_{t+h})^2 | S_t] \Rightarrow F_t^* x_{t+h}(S_t) = \mathbb{E}[x_{t+h}|S_t]. \tag{B.1}$$

It immediately follows that the loss from an imprecise forecast is the variance of x_{t+h} conditional on S_t

$$\mathbb{E}[(F_t^* x_{t+h}(S_t) - x_{t+h})^2 | S_t] = \text{var}(x_{t+h}|S_t). \tag{B.2}$$

Moreover, we can decompose this variance in terms of uncertainty about the long-run mean and variance of short-run fluctuations:

$$\text{var}(x_{t+h}|S_t) = \text{var}((1-\rho^h)\mu + \rho^h x_t + \sum_{j=1}^{h} \rho^{h-j} \varepsilon_{t+j} | S_t) \tag{B.3}$$

$$= (1-\rho^h)^2 \text{var}(\mu|S_t) + \sigma_\varepsilon^2 \sum_{j=1}^{h} \rho^{2(h-j)}, \tag{B.4}$$

where the second line follows from:

1. orthogonality of future innovations to S_t that follows from feasibility ($\varepsilon_{t+j} \perp \mathcal{A}_t, \forall j \geq 1$);
2. $\text{var}(x_t|S_t) = 0$ since $x_t \in S_t$ by assumption.

It is important to note that the second term in Equation B.4 is independent of the choice for S_t. We can now rewrite the agent's problem as:

$$\min_{S_t} \mathbb{E}[(1-\rho^h)^2 \text{var}(\mu|S_t) + C(S_t)|x_t] \tag{B.5}$$

$$s.t.\ \{x_t\} \subseteq S_t \subseteq \mathcal{A}_t, \tag{B.6}$$

where the expectation $\mathbb{E}[.|x_t]$ is taken conditional on x_t because the choice for what information to utilize happens after the agent observes the context but before information is processed.

The next step in the proof is to show that under the optimal information utilization, the distribution of $\mu|S_t$ is Gaussian. To prove this, we show that for any arbitrary $S_t \in \mathcal{A}_t$, there exists another $\hat{S}_t \in \mathcal{A}_t$ that (1) induces a Gaussian posterior and (2) yields a lower value for the objective function than S_t. To see this, let $S_t \supseteq \{x_t\}$ be in $\mathcal{A}_t$ and let $\hat{S}_t \supseteq \{x_t\}$ be such that

$$\text{var}(\mu|\hat{S}_t) = \mathbb{E}[\text{var}(\mu|S_t)|x_t].$$

Such a $\hat{S}_t$ exists because $\mathcal{A}_t$ is assumed to contain all possible signals on μ that are feasible, so if an expected variance is attainable under an arbitrary signal, it is also attainable by a Gaussian signal. Since both signals imply the same expected variance, to prove our claim, we only need to show that $C(\hat{S}_t) \leq C(S_t)$. To see this, recall that $C(S_t)$ is monotonically increasing in $\mathbb{I}(S_t, x_{t+h}|x_t)$. Thus,

$$C(\hat{S}_t) \leq C(S_t) \Leftrightarrow \mathbb{I}(\hat{S}_t, x_{t+h}|x_t) \leq \mathbb{I}(S_t, x_{t+h}|x_t). \tag{B.7}$$

A final observation yields our desired result: by definition of the mutual information function in terms of entropy,[35]

$$\mathbb{I}(S_t; \mu|x_t) = h(\mu|x_t) - \mathbb{E}[h(\mu|S_t)|x_t]. \tag{B.8}$$

Similarly,

$$\mathbb{I}(\hat{S}_t; \mu|x_t) = h(\mu|x_t) - \mathbb{E}[h(\mu|\hat{S}_t)|x_t]. \tag{B.9}$$

It follows from these two observations that

$$C(\hat{S}_t) \leq C(S_t) \Leftrightarrow \mathbb{E}[h(\mu|\hat{S}_t)|x_t] \geq \mathbb{E}[h(\mu|S_t)|x_t]. \tag{B.10}$$

The right hand side of this condition is true by the maximum entropy of Gaussian random variables among random variables with the same variance, with equality only if both S_t and $\hat{S}_t$ are Guassian (see for example Cover and Thomas (1991)).[36] This result implies that $C(\hat{S}_t) \leq C(S_t)$. Therefore, for any arbitrary $S_t \subset \mathcal{A}_t$ such that $\mu|S_t$ is non-Gaussian, we have shown that there exists $\hat{S}_t \subset \mathcal{A}_t$ that is (1) feasible and (2) strictly preferred to S_t and (3) $\mu|\hat{S}_t$ is Gaussian.

Hence, without loss of generality, we can assume that under the optimal retrieval of information, $\mu|S_t$ is normally distributed. Now, for a Gaussian $\{x_t\} \subset S_t \subset \mathcal{A}_t$, since entropy of Gaussian random variables are linear in the log of their variance, we have:

$$\mathbb{I}(\mu; S_t|x_t) = h(\mu|x_t) - h(\mu|S_t) \tag{B.11}$$

$$= \frac{1}{2\ln(2)} \ln(\text{var}(\mu|x_t)) - \frac{1}{2\ln(2)} \ln(\text{var}(x_t|S_t)). \tag{B.12}$$

Moreover, for simplicity we define $\tau(S_t) \equiv \text{var}(\mu|S_t)^{-1}$ as the precision of belief about μ generated by S_t and $\underline{\tau} \equiv \text{var}(\mu|x_t)^{-1}$ as the precision of the prior belief of the agent about μ. Moreover, for ease of notation and wihtout loss of generality, we normalize the mutual information function by the constant $2\ln(2)$.[37] It follows

[35]For random variables (X, Y), $\mathbb{I}(X; Y) = h(X) - \mathbb{E}^Y[h(X|Y)]$ where for any random variable Z with PDF $f_Z(z)$, $h(Z)$ is the entropy of Z defined as the expectation of negative log of its PDF: $h(Z) = -\mathbb{E}^Z[\log_2(f_Z(Z))]$.

[36]For completeness, we briefly outline the proof for maximum entropy of Gaussian random variables. The claim is: among all the random variables X variance σ^2, X has the highest entropy if it is normally distributed. The proof follows from optimizing over the PDF of the distribution of X:

$$\max_{\{f(x)\geq 0: x\in\mathbb{R}\}} -\int_{x\in\mathbb{R}} f(x)\log(f(x))dx \tag{maximum entropy}$$

$$s.t. \int_{x\in\mathbb{R}} x^2 f(x)dx - (\int_{x\in\mathbb{R}} xf(x)dx)^2 dx = \sigma^2 \tag{constraint on variance}$$

$$\int_{x\in\mathbb{R}} f(x)dx = 1. \tag{constraint on f being a PDF}$$

[37]Alternatively, one can re-scale both γ and ω by $2\ln(2)$, which is simply a normalization of their values.

that

$$\mathbb{I}(\mu; S_t|x_t) = \ln\left(\frac{\tau(S_t)}{\underline{\tau}}\right), \tag{B.13}$$

$$C(S_t) = \omega\frac{\exp(\gamma\cdot\mathbb{I}(\mu; S_t|x_t)) - 1}{\gamma} \tag{B.14}$$

$$= \omega\frac{\left(\frac{\tau(S_t)}{\underline{\tau}}\right)^{\gamma} - 1}{\gamma}. \tag{B.15}$$

Hence, the agent's problem can be rewritten as

$$\min_{S_t}\mathbb{E}\left[\frac{(1-\rho^h)^2}{\tau(S_t)} + \omega\frac{\left(\frac{\tau(S_t)}{\underline{\tau}}\right)^{\gamma} - 1}{\gamma}\Big| x_t\right] \tag{B.16}$$

$$s.t.\ \{x_t\} \subseteq S_t \subseteq \mathcal{A}_t. \tag{B.17}$$

Finally, since the objective of the agent only depends on the precision induced by S_t, we can reduce the problem to directly choosing this precision, where the constraint on S_t implies bounds on achievable precision: the precision should be bounded below by $\underline{\tau}$ (since the agent knows x_t). Moreover, it has to be bounded above by $\text{var}(\mu|x^t)^{-1}$ which is the precision after utilizing *all available information*. Replacing these in the objective, and changing the choice variable to $\tau(S_t)$ we arrive at the exposition delivered in the lemma. □

B.3 Proposition 1

Proof. We start by solving the simplified problem in Lemma 1. The problem has two constraints for τ: $\tau \geq \underline{\tau}$ and $\tau \leq \bar{\tau}_t \equiv var(\mu|x^t)^{-1}$. By assumption $\text{var}(\mu|x^t)$ is arbitrarily small so we can assume that the second constraint does not bind. The K-T conditions with respect to τ are

$$-\frac{(1-\rho^h)^2}{\tau^2} + \frac{\omega}{\tau}\left(\frac{\tau}{\underline{\tau}}\right)^{\gamma} \geq 0, \quad \tau \geq \underline{\tau}, \quad \left(-\frac{(1-\rho^h)^2}{\tau^2} + \frac{\omega}{\tau}\left(\frac{\tau}{\underline{\tau}}\right)^{\gamma}\right)(\tau - \underline{\tau}) = 0.$$

Therefore, the variance of the agent's belief about the long-run mean is given by

$$\text{var}(\mu|S_t) = \tau^{-1} = \underline{\tau}^{-1}\min\left\{1, \left(\frac{\omega\underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}\right\}. \tag{B.18}$$

The next step is to find an optimal signal set $S_t \supseteq \{x_t\}$ that generates this posterior, so we can characterize how the agent's forecast correlates with the recent observation. In particular, the regression considered in the Proposition (and more generally in our analysis) is interested in identifying the conditional mean of the agent's forecast ($F_t x_{t+h}$) given the recent observation x_t and the true mean μ, which we denote for the rest of the proof as $\mu_t \equiv \mathbb{E}[F_t x_{t+h}|x_t, \mu]$. Two cases arise:

1. if $\left(\frac{\omega\underline{\tau}}{(1-\rho^h)^2}\right) \geq 1$, then $\sigma^2 = (1-\rho^h)^2\underline{\tau}$ and $S_t = \{x_t\}$ delivers us the agent's posterior. In other words, $\text{var}(\mu|S_t) = \text{var}(\mu|x_t)$ meaning that the agents does not retrieve any further information other than what is implied by the context. In this case, $\mathbb{E}[\mu|S_t] = \mathbb{E}[\mu|x_t] = x_t$ and

$$\mu_t \equiv \mathbb{E}[\mathbb{E}[x_{t+h}|S_t]|\mu, x_t] = (1-\rho^h)\mathbb{E}[\mathbb{E}[\mu|S_t]|\mu, x_t] + \rho^h\mathbb{E}[\mathbb{E}[x_t|S_t]|\mu, x_t] = x_t \tag{B.19}$$

2. if $\left(\frac{\omega\underline{\tau}}{(1-\rho^h)^2}\right) < 1$, then it means that the agent utilizes more information than what is revealed by the

context x_t. Suppose a signal structure $\tilde{S}_t$ generates this posterior variance.[38] By Lemma 1 this has to be Guassian. First, it is convenient to observe that the set $\hat{S}_t \equiv \{x_t, \mathbb{E}[\mu|\tilde{S}_t]\}$ is a sufficient statistic for $\tilde{S}_t$. To see the equivalence of the two sets, note that both are comprised of Gaussian variables and by the law of total variance, both sets generate the same posterior variance for the agent.[39]

Now, by Bayesian updating for Gaussians:

$$\mathbb{E}[\mu|S_t] = \mathbb{E}[\mu|\tilde{S}_t] = \mathbb{E}[\mu|x_t] + \frac{\operatorname{cov}(\mu, \mathbb{E}[\mu|\tilde{S}_t]|x_t)}{\operatorname{var}(\mathbb{E}[\mu|\tilde{S}_t]|x_t)}(\mathbb{E}[\mu|\tilde{S}_t] - \mathbb{E}[\mu|x_t]).$$

Since $\mathbb{E}[\mu|\tilde{S}_t] - \mathbb{E}[\mu|x_t] \neq 0$ almost surely, this implies that

$$\operatorname{cov}(\mu, \mathbb{E}[\mu|\tilde{S}_t]|x_t) = \operatorname{var}(\mathbb{E}[\mu|\tilde{S}_t]|x_t) = \underline{\tau}^{-1} - \tau^{-1}, \tag{B.20}$$

where the last equality follows from the law of total variance. Now, consider the following decomposition of $\mathbb{E}[\mu|\tilde{S}_t]$:

$$\mathbb{E}[\mu|\tilde{S}_t] = a\mu + bx_t + \varepsilon_t,$$

where a and b are constants and ε_t is the residual that is orthogonal to both x_t and μ conditional on $\tilde{S}_t$. By agent's prior on μ, we have

$$x_t = \mathbb{E}[\mu|x_t] = \mathbb{E}[\mathbb{E}[\mu|\tilde{S}_t]|x_t] = a\mathbb{E}[\mu|x_t] + bx_t = (a+b)x_t,$$

so $a + b = 1$. Moreover, we also have

$$\operatorname{cov}(\mu, \mathbb{E}[\mu|\tilde{S}_t]|x_t) = a\operatorname{var}(\mu|x_t),$$

so $a = 1 - \frac{\underline{\tau}}{\tau}$. Therefore,

$$\begin{aligned} &\mathbb{E}[\mathbb{E}[\mu|\tilde{S}_t]|\mu, x_t] = (1 - \frac{\underline{\tau}}{\tau})\mu + \frac{\underline{\tau}}{\tau}x_t \\ \Rightarrow &\mu_t \equiv \mathbb{E}[\mathbb{E}[x_{t+h}|\tilde{S}_t]|\mu, x_t] = (1-\rho^h)(1 - \frac{\underline{\tau}}{\tau})\mu + (1-\rho^h)\frac{\underline{\tau}}{\tau}x_t + \rho^h x_t. \end{aligned} \tag{B.21}$$

Now subtracting the fully informed rational forecast $E_t[x_{t+h}] \equiv (1-\rho^h)\mu + \rho^h x_t$, we have

$$\mu_t = E_t x_{t+h} + (1-\rho^h)\frac{\underline{\tau}}{\tau}(x_t - \mu) \tag{B.22}$$

Note that this case also subsumes the previous case, because if $S_t = x_t$ then $\underline{\tau} = \tau$ and $\mu_t = x_t$ as before.

[38] It is important to note that the model does not discriminate on which observations are in S_t but only the quantity of information revealed by those observations because the agent's payoff depends on the variance of her forecasts

[39] Formally, we have

$$\begin{aligned} \operatorname{var}(\mu|x_t) &= \operatorname{var}(\mu|\tilde{S}_t) + \operatorname{var}(\mathbb{E}[\mu|\tilde{S}_t]|x_t) \\ \operatorname{var}(\mu|x_t) &= \operatorname{var}(\mu|\hat{S}_t) + \operatorname{var}(\mathbb{E}[\mu|\hat{S}_t]|x_t), \end{aligned}$$

but note that $\operatorname{var}(\mathbb{E}[\mu|\hat{S}_t]|x_t) = \operatorname{var}(\mathbb{E}[\mu|x_t, \mathbb{E}[\mu|\tilde{S}_t]]|x_t) = \operatorname{var}(\mathbb{E}[\mu|\tilde{S}_t]|x_t)$. Thus, it has to be that $\operatorname{var}(\mu|\tilde{S}_t) = \operatorname{var}(\mu|\hat{S}_t)$

Finally, define $u_t \equiv F_t x_{t+h} - \mu_t$. Plugging in the expression for τ from (B.18) into (B.22), and setting (normalizing) $\mu = 0$, we get the expression of interest:

$$F_t x_{t+h} = \mu_t + u_t = E_t x_{t+h} + (1-\rho^h) \min\left\{1, \left(\frac{\omega \underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}\right\} x_t + u_t \tag{B.23}$$

where $\mathbb{E}[u_t | x_t, \mu] = 0$. □

B.4 Proposition 2

Proof. From Proposition 1 we can derive Δ as

$$\Delta = (1-\rho^h) \min\left\{1, \left(\frac{\omega \underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}\right\}. \tag{B.24}$$

1. Note that if $\Delta = 0$ then either $\rho^h = 1$ or $\omega = 0$, but recall that this expression for the precision of the long-run mean was derived under the assumption that $\text{var}(\mu | x^t)$ is arbitrarily small. So $\Delta = 0$ if and only if either $\rho = 1$ or $\omega = 0$ and past information potentially available to the forecaster is infinite.
2. As long as $\gamma \geq 0$, which is true by assumption, it is straightforward to verify that Δ is increasing in ω and $\underline{\tau}$.
3. For Δ to be decreasing in ρ^h it has to be the case that $(1-\rho^h)^{1-\frac{2}{1+\gamma}}$ is decreasing in ρ^h, which is the case if and only if

$$1 - \frac{2}{1+\gamma} \geq 0 \Leftrightarrow \gamma \geq 1. \tag{B.25}$$

4. We then prove the comparative static results for $\zeta(\rho, h)$. From Proposition 2 we have

$$\ln(\zeta(\rho, h)) = \ln\left(1 + (\rho^{-h} - 1) \min\left\{1, \left(\frac{\omega \underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}\right\}\right). \tag{B.26}$$

   It is straightforward to see that the term inside the log on the right hand side is larger than 1, so the implied persistence is larger than the actual persistence. Moreover, for $\zeta(\rho, h)$ to be decreasing in ρ^h, $(1-\rho^h)^{1-\frac{2}{1+\gamma}} / \rho^h$ needs to be decreasing in ρ^h, which is true if and only if $\gamma \geq 2\rho^h - 1$. Therefore, for ζ to be decreasing for any value of ρ^h, we need $\gamma \geq 1$.

□

B.5 Corollary 1

Proof. First, we prove the comparative static with respect to ρ. The statement holds trivially when $\rho_h^s = 1$, which happens when the minimum in the expression for Δ binds. Thus, it suffices to consider the case where $1 > \left(\frac{\omega \underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}$. Then, denoting $f(\rho^h) = (\rho_h^s)^h = \rho^h + (1-\rho^h) \min\left\{1, \left(\frac{\omega \underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}\right\}$, f is differentiable in the region of interest. Given that $\Delta = f(\rho^h) - \rho^h$ is decreasing in ρ^h by Proposition 2, $f'(\rho^h) < 1$.

Then, the comparative static with respect to ρ holds if:

$$\frac{\partial \rho_h^s - \rho}{\partial \rho} = \rho^{h-1} \cdot f'(\rho^h) \cdot f(\rho^h)^{\frac{1}{h}-1} - 1 < 0$$
$$\iff \rho^{h-1} f(\rho^h)^{\frac{1}{h}-1} \cdot f'(\rho^h) < 1$$

Given that $f' < 1$, it suffices to show that $\rho^{h-1} \cdot f(\rho^h)^{\frac{1}{h}-1} < 1 \iff \rho^{h-1} < f(\rho^h)^{\frac{h-1}{h}} \iff \rho^h f(\rho^h)$, which holds trivially from the definition.

Second, to compute the comparative static with respect to h, we examine two cases. First, if $1 \leq \left(\frac{\omega\underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}} \iff \rho^h > 1 - \sqrt{\omega\underline{\tau}}$. In this case, $\rho_h^s = 1$, so our result holds naturally. Thus, it suffices to only consider the case $\rho^h < 1 - \sqrt{\omega\underline{\tau}}$. Note that if $\omega\underline{\tau} \geq 1$, this region is the empty set for positive ρ and h, and the statement is trivially true for such parameters. Thus, we only need to consider the cases where $\omega\underline{\tau} < 1$ where the interval $[0, 1 - \sqrt{\omega\underline{\tau}}]$ has positive measure. In this case,

$$\frac{\partial \rho_h^s}{\partial h} = -\frac{1}{h^2} \log f + \frac{1}{h^2} \frac{f'(\rho^h)}{f(\rho^h)} \cdot \log(\rho^h) \cdot \rho^h.$$

Consequently, $\frac{\partial \rho_h^s}{\partial h} \geq 0$ if

$$\psi(\rho^h) = -\log f(\rho^h) + \frac{f'(\rho^h)}{f(\rho^h)} \log \rho^h \cdot \rho^h \geq 0$$

It is easy to see that for $\psi(\rho^h)$ is continuous and well-defined in the region of interest ($\rho^h \in [0, 1 - \sqrt{\omega\tau}]$). Also, note that $\lim_{\rho^h \mapsto 0} \psi(\rho^h) = (\omega\underline{\tau})^{\frac{1}{1+\gamma}} > 0$ and $\lim_{\rho^h \mapsto 1-\sqrt{\omega\tau}} \psi(\rho^h) = \frac{2(1-\sqrt{\omega\tau})}{1+\gamma} \log(1 - \sqrt{\omega\tau}) < 0$. Consequently, by the intermediate value theorem, there exists a $\lambda^* > 0$ such that for $\rho^h \in [0, \lambda^*]$, $\psi(\rho^h) \geq 0$ and thus ρ_h^s is increasing in h, where λ^* is *independent* of ρ and h. Consequently, for any $\rho < 1$, there exists an $h^*(\rho) = \log(\lambda^*)/\log(\lambda)$ such that $\rho_h^s - \rho$ is increasing in h for $h \geq h^*(\rho)$. □

C Generalized Model for ARMA Processes

We consider a Markov Gaussian process $\{X_t : t \geq 0\}$ on $\mathbb{R}^n$ with the following state space representation:

$$X_t = (I - A)\bar{X} + AX_{t-1} + Qu_t.$$

Suppose the agent's task is to make a set of forecasts of horizon h_i for a vector of m variables $Y_t = (y_{i,t+h_i})_{i \in \{1,\ldots,m\}}$, where $y_{i,t+h_i} = w_i' X_{t+h_i}$ is a linear combination of X_{t+h_i}. Since innovations u_t are i.i.d. over time, the agent's forecast of X_{t+h} for any $h \geq 0$ at a given time t can be written as

$$E[X_{t+h}|S_t] = (I - A^h)E[\bar{X}|S_t] + A^h X_t,$$

where S_t is what is on top of the agent's mind at time t. Thus, for any $y_{i,t+h_i}$:

$$E[y_{i,t+h_i}|S_t] = w_i'(I - A^h)E[\bar{X}|S_t] + w_i' A^h X_t.$$

Assuming that the agent minimizes a squared sums of errors weighted by W, the resulting objective can be written as

$$\begin{aligned}&-\frac{1}{2}E[(Y_t - E[Y_t|S_t])'W(Y_t - E[Y_t|S_t])|S_t]\\&= -\frac{1}{2}tr(\Sigma_t HWH') + \text{terms independent of optimization},\end{aligned}$$

where $\Sigma_t = Var(\bar{X}|S_t)$ is the variance of the long-run mean of X_t given S_t and H is an $n \times m$ matrix whose j'th column is $(I - A^h)'w_i$. We define $\Omega \equiv HWH'$. Then, the agent's loss at time t from not knowing the long-run mean is given by $-\frac{1}{2}tr(\Sigma_t\Omega)$.

Suppose now that the agent's prior at the beginning of the period is $\bar{X}|X_t \sim N(X_t, \underline{\Sigma})$, which is a generalized version of the prior assumed in the main text. Conditional on this prior, the agent solves the following problem (the derivations for which closely follow the proof of Lemma 1):

$$\begin{aligned}&\max_{\Sigma}\left\{-tr(\Omega\Sigma) - \omega\frac{(|\underline{\Sigma}||\Sigma|^{-1})^{\gamma} - 1}{\gamma}\right\}\\&s.t. \mathbf{0} \preceq \Sigma \preceq \underline{\Sigma},\end{aligned}$$

where $(\succeq 0)$ denotes positive-semidefiniteness. This is a convex optimization problem on the *positive semi-definite cone*, similar to the problem studied in Afrouzi and Yang (2020). While Afrouzi and Yang (2020) only consider the case of $\gamma \to 0$, we solve for the more general case of $\gamma > 0$. Since the cost of inaccuracy approaches infinity if $|\Sigma| \to 0$, the optimal subjective variance Σ should have a strictly positive determinant, with all the eigenvalues of Σ strictly positive ($\Sigma \succ \mathbf{0}$). In other words, we can ignore the constraint $\Sigma \succ 0$ as it should not bind under the solution. On the other hand, the constraint $\Sigma \preceq \underline{\Sigma}$, however, potentially binds and needs to be considered (this intuitively corresponds to the case in which zero costly learning occurs).

We assume Λ is the generalized Lagrange multiplier on this constraint. It follows from convex optimization that Λ is also positive semi-definite, commutes with $X \equiv \underline{\Sigma} - \Sigma$, and satisfies complementarity slackness $\Lambda X = X\Lambda = \mathbf{0}$ (See Afrouzi and Yang (2020) for details). The first order condition is then

$$\Omega = \omega|\underline{\Sigma}|^{\gamma}|\Sigma|^{-\gamma}\Sigma^{-1} + \Lambda,$$

which can be rewritten as

$$\Omega X = \Omega\underline{\Sigma} - \omega|\underline{\Sigma}|^{\gamma}|\Sigma|^{-\gamma} + \Lambda\underline{\Sigma}.$$

Now multiply this by $\underline{\Sigma}^{\frac{1}{2}}$ from left and $\underline{\Sigma}^{-\frac{1}{2}}$ from the right, and observe that

$$\underline{\Sigma}^{\frac{1}{2}}\Omega\underline{\Sigma}^{\frac{1}{2}}\underline{\Sigma}^{-\frac{1}{2}}X\underline{\Sigma}^{-\frac{1}{2}} = \underline{\Sigma}^{\frac{1}{2}}\Omega\underline{\Sigma}^{\frac{1}{2}} - \omega|\underline{\Sigma}|^{\gamma}|\Sigma|^{-\gamma}I + \underline{\Sigma}^{\frac{1}{2}}\Lambda\underline{\Sigma}^{\frac{1}{2}}.$$

Setting $\hat{\Omega} = \underline{\Sigma}^{\frac{1}{2}}\Omega\underline{\Sigma}^{\frac{1}{2}}$, $\hat{X} = \underline{\Sigma}^{-\frac{1}{2}}X\underline{\Sigma}^{-\frac{1}{2}}$, $\hat{\Lambda} = \underline{\Sigma}^{\frac{1}{2}}\Lambda\underline{\Sigma}^{\frac{1}{2}}$, and $\hat{\omega} = \omega|\underline{\Sigma}|^{\gamma}|\Sigma|^{-\gamma}$, we obtain:

$$\hat{\Omega}\hat{X} = \hat{\Omega} - \hat{\omega}I + \hat{\Lambda}. \tag{C.1}$$

Note that $\hat{X}\hat{\Lambda} = \hat{\Lambda}\hat{X} = 0$. We can also see that $\hat{\Omega}\hat{X} = \hat{X}\hat{\Omega}$ since the right hand side of Equation (C.1) above is symmetric. Finally, we can see that $\hat{\Lambda}$ and $\hat{\Omega}$ also commute.[40] Thus, since $\hat{\Omega}$, $\hat{X}_t$ and $\hat{\Lambda}$ are all symmetric, they are all diagonalizable, and given that they all commute with one another, they must be simultaneously diagonalizable. This implies that there are diagonal matrices D_{Λ}, D_X and D_{Ω}, as well as an orthonormal basis U ($UU' = U'U = I$), such that

$$\hat{\Omega} = UD_{\Omega}U',\ \hat{X} = UD_XU',\ \hat{\Lambda} = UD_{\Lambda}U'$$

[40]To see this, multiply the Equation (C.1) by $\hat{\Lambda}$ form right and note that $\hat{\Omega}\hat{\Lambda}$ has to be symmetric, indicating that $\hat{\Lambda}\hat{\Omega} = (\hat{\Lambda}\hat{\Omega})' = \hat{\Omega}\hat{\Lambda}$.

Now multiplying Equation (C.1) by U from left and U' from right, we have

$$D_\Omega D_X = D_\Omega - \hat{\omega} I + D_\Lambda, \quad D_\Lambda \succeq 0, \quad D_X \succeq 0, \quad D_X D_\Lambda = 0.$$

Given that these equations are in terms of diagonal matrices, the inequality needs to hold entry-by-entry on the diagonal, implying that for any $1 \leq i \leq n$:

$$D_{X,ii} = 1 - \hat{\omega} \max\{D_{\Omega,ii}, \hat{\omega}\}^{-1},$$

or in matrix form:

$$I - \hat{X} = \max\{\frac{\hat{\Omega}}{\hat{\omega}}, I\}^{-1} = \max\{\frac{\underline{\Sigma}^{\frac{1}{2}} \Omega \underline{\Sigma}^{\frac{1}{2}}}{\hat{\omega}}, I\}^{-1}, \tag{C.2}$$

or

$$\Sigma = \underline{\Sigma}^{\frac{1}{2}} \max\{\frac{\underline{\Sigma}^{\frac{1}{2}} \Omega \underline{\Sigma}^{\frac{1}{2}}}{\hat{\omega}}, I\}^{-1} \underline{\Sigma}^{\frac{1}{2}}, \tag{C.3}$$

where the only unknown on right hand side is $\hat{\omega}$.

To calculate $\hat{\omega}$, take the determinant of the above equation and note that

$$\det(I - \hat{X}) = \det(I - \underline{\Sigma}^{-\frac{1}{2}} X \underline{\Sigma}^{-\frac{1}{2}}) = \det(\underline{\Sigma}^{-1}\Sigma) = (\frac{\hat{\omega}}{\omega})^{-\gamma^{-1}}.$$

Thus, taking the log-determinant of Equation (C.2) (which is permitted because both sides are strictly positive definite) gives:

$$\log(\hat{\omega}) = \log(\omega) + \gamma \log\det\left(\max\{\frac{\underline{\Sigma}^{\frac{1}{2}} \Omega \underline{\Sigma}^{\frac{1}{2}}}{\hat{\omega}}, I\}\right).$$

Now let $\{\lambda_i\}_{i \in \{1,\ldots,n\}}$ denote the eigenvalues of the matrix $\underline{\Sigma}^{\frac{1}{2}} \Omega \underline{\Sigma}^{\frac{1}{2}}$ (note that these are simply parameters of the model). Then, we can rewrite this equation as

$$\log(\hat{\omega}) = \log(\omega) + \gamma \sum_{\lambda_i \geq \hat{\omega}} \log(\frac{\lambda_i}{\hat{\omega}}). \tag{C.4}$$

which is an equation only in terms of $\hat{\omega}$ and unique to our case.

To prove the existence of a solution, note that the left hand side is increasing in $\hat{\omega}$ and subjects onto all of $\mathbb{R}$. On the other hand, the right hand side is decreasing in $\hat{\omega}$, with its range being $[\log(\omega), \infty)$. Thus, there is a unique $\hat{\omega}$ that solves this equation (which incidentally is larger than ω for $\gamma > 0$ as long as there is at least one eigenvalue larger than ω). Thus Equations (C.3) and (C.4) together pin down the optimal Σ for the agent. Therefore, applying standard Kalman filtering results, we obtain that the agent's belief about the long run mean is given by

$$\bar{X} | S_t \sim N(\hat{X}_t, \Sigma),$$

where

$$E[\hat{X}_t | \bar{X}, X_t] = \bar{X} + \underbrace{\underline{\Sigma}^{\frac{1}{2}} \max\{\frac{\underline{\Sigma}^{\frac{1}{2}} \Omega \underline{\Sigma}^{\frac{1}{2}}}{\hat{\omega}}, I\}^{-1} \underline{\Sigma}^{-\frac{1}{2}} (X_t - \bar{X})}_{\text{overreaction}}.$$

and Σ is the solution in Equation (C.3).

Consequently, as is the case for our simple AR(1) example, there is a positive loading on the subjective long-run mean on the most recent observation, which yields overreaction.

D Underreaction

Our model can be extended in a simple way to accommodate underreaction. Following the noisy information literature (e.g. Woodford (2003) and Khaw, Li and Woodford (2018)), we now assume that the individual receives a noisy signal of x_t:

$$s_t = x_t + \epsilon_t, \epsilon_t \sim N\left(0, \tau_\epsilon^{-1}\right). \tag{D.1}$$

Furthermore, the agent has a prior over the latent value x_t, given by $x_t \sim N\left(\bar{x}, \tau_0^{-1}\right)$. In this case, the agent obtains the posterior beliefs regarding the most recent signal:

$$\hat{x}_t | s_t = \frac{\tau_\epsilon}{\tau_0 + \tau_\epsilon} s_t + \frac{\tau_0}{\tau_0 + \tau_\epsilon} \bar{x}. \tag{D.2}$$

We do not need to take a stance on $\bar{x}$: as long as the prior does not depend on the value of x_t, all of our conclusions are unchanged. The agent then forms a default belief regarding the long-run mean μ centered around the noisy recent signal $\hat{x}_t$:

$$\hat{\mu} \sim N\left(\hat{x}_t, \underline{\tau}\right). \tag{D.3}$$

Our main model can be seen as a special case ($\tau_\epsilon \mapsto \infty$) of this more general case that allows for noisy signals.

The derivations are similar as before and we have:

$$E[\mu | \hat{x}_t, S_t] = \min\left\{1, \left(\frac{\omega \underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}\right\} \hat{x}_t \tag{D.4}$$

$$F_t x_{t+h} = \rho^h \cdot \hat{x}_t + (1-\rho^h) \min\left\{1, \left(\frac{\omega \underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}\right\} \hat{x}_t + \underbrace{\varepsilon_t}_{\text{noise}} \tag{D.5}$$

$$= \rho^h x_t + \left[\underbrace{\frac{\tau_\epsilon}{\tau_0 + \tau_\epsilon}(1-\rho^h) \min\left\{1, \left(\frac{\omega \underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}\right\}}_{\text{overreaction}} - \underbrace{\frac{\tau_0}{\tau_0 + \tau_\epsilon}\rho^h}_{\text{underreaction}}\right] x_t + \text{constant} + \varepsilon_t. \tag{D.6}$$

Note that when $\tau_\epsilon \mapsto \infty$, the equation above converges to our expression in the main text. However, for finite τ_ϵ, noisy signals introduce a downward pressure on the loading of the forecast on x_t, which counteracts overreaction. The intuition is simple: the agent's forecast overreacts to $\hat{x}_t$, but with noisy information, $\hat{x}_t$ itself underreacts to x_t. The following proposition derives the conditions for when each force dominates. When the noise in the signal is small, overreaction is the dominant force.

The above expression implies the following proposition, which shows that in this model extension the degree of overreaction is still stronger when the process is less persistent (i.e., ρ is small):

Proposition D.1. Holding fixed the noisy information parameters $\tau_\epsilon, \tau_0 < \infty$, there is overreaction ($\rho_h^s > \rho$) for sufficiently low ρ, and underreaction ($\rho_h^s < \rho$) if $\rho \mapsto 1$. If $\gamma \geq 1$, $\Delta = \rho_{s,h} - \rho^h$ is decreasing in ρ^h.

Proof. We have:

$$\rho_{s,h} - \rho^h = \frac{\tau_\epsilon}{\tau_0 + \tau_\epsilon}(1-\rho^h) \min\left\{1, \left(\frac{\omega \underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}\right\} - \frac{\tau_0}{\tau_0 + \tau_\epsilon}\rho^h. \tag{D.7}$$

It is evident that the expression on the right hand side is positive as $\rho \mapsto 0$ (it converges to $\frac{\tau_\epsilon}{\tau_0+\tau_\epsilon}\left(\omega\underline{\tau}\right)^{\frac{1}{1+\gamma}}$), and negative as $\rho \mapsto 1$ (it converges to $-\frac{\tau_0}{\tau_0+\tau_\epsilon}$). For intermediate values of ρ, when ρ is sufficiently high such that $\frac{\omega\underline{\tau}}{(1-\rho^h)^2} > 1$, the right hand side becomes:

$$\frac{\tau_\epsilon}{\tau_0 + \tau_\epsilon} - \rho^h, \tag{D.8}$$

which is monotonically decreasing in ρ. When ρ is sufficiently low such that $\frac{\omega\underline{\tau}}{(1-\rho^h)^2} < 1$, the expression becomes:

$$\frac{\tau_\epsilon}{\tau_0 + \tau_\epsilon}(1-\rho^h)\left(\frac{\omega\underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}} - \frac{\tau_0}{\tau_0 + \tau_\epsilon}\rho^h = \frac{\tau_\epsilon}{\tau_0 + \tau_\epsilon}(\omega\underline{\tau})^{\frac{1}{1+\gamma}}(1-\rho^h)^{-\frac{\gamma-1}{1+\gamma}} - \frac{\tau_0}{\tau_0 + \tau_\epsilon}\rho^h. \tag{D.9}$$

If we assume $\gamma \geq 1$, each of the terms are decreasing in ρ^h, which is in line with the empirical evidence.

Overall, in our experiment, the signals are rather simple and unambiguous, so the noise is likely very small. In other environments, signals can be noisier, which may generate underreaction even at the individual level. Similarly, if we introduce in our model frictions such as insufficient attention and infrequent updating (Mankiw and Reis, 2002), then we can also obtain underreaction. This is unlikely to be the case in our experiment, but it could be more relevant for other settings such as households' expectations of inflation. □

E Model Predictions for Changing What's on Top of Mind

In this section, we describe our model's predictions for the additional experiments in Section 5.5 (where we change what's on top of mind).

E.1 Setup

We have two main experimental designs to change what is on top of mind for participants. In the first condition, we show a red line corresponding to $x = 0$. In the second condition, we require participants to click on x_{t-10} in each round before they can make new forecasts. Both designs aim to change the default context from the original default, i.e., the most recent realization x_t.

In our baseline model, prior beliefs are given by a normal distribution with mean x_t and precision $\underline{\tau}$. We model these additional tests as providing an extra signal of the long-run mean, I, before the agent decides what information to utilize. By design, this signal is on average centered around 0 with precision $\bar{\tau}'$. After seeing the signal I, the belief the agent has regarding the long-run mean is given by:

$$\mu | x_t, I \sim N(z_t, \underline{\tau} + \bar{\tau}') \tag{E.1}$$

Standard Gaussian updating implies that $E[z_t|x_t] = \alpha x_t$, where $\alpha = \frac{\underline{\tau}}{\underline{\tau}+\bar{\tau}'} < 1$.

After processing the signal, the agent then processes additional information. Following our experimental design, we assume $h = 1$ for simplicity. Using the same computation as in the main model, we obtain:

$$E[\mu|x_t, S_t, I] = \min\left\{1, \left(\frac{\omega(\underline{\tau} + \bar{\tau}')}{(1-\rho)^2}\right)^{\frac{1}{1+\gamma}}\right\} z_t, \tag{E.2}$$

and consequently:

$$\rho^s_{1,I} = \rho + (1-\rho)\cdot \min\left\{1, \left(\frac{\omega(\underline{\tau} + \bar{\tau}')}{(1-\rho)^2}\right)^{\frac{1}{1+\gamma}}\right\} \cdot \frac{\underline{\tau}}{\underline{\tau} + \bar{\tau}'}. \tag{E.3}$$

In comparison, our original expression is:

$$\rho_1^s = \rho + (1-\rho)\cdot \min\left\{1, \left(\frac{\omega\underline{\tau}}{(1-\rho)^2}\right)^{\frac{1}{1+\gamma}}\right\}. \tag{E.4}$$

E.2 Result

We have the following proposition:

Proposition E.1. The implied persistence curve in the new conditions $\rho_{1,I}^s$ lies below the original implied persistence curve ρ_1^s. In other words, $\rho_{1,I}^s < \rho_1^s$ for each level of actual ρ (except $\rho = 1$).

Proof. It suffices to show:

$$\min\left\{1, \left(\frac{\omega\underline{\tau}}{(1-\rho)^2}\right)^{\frac{1}{1+\gamma}}\right\} > \min\left\{1, \left(\frac{\omega(\underline{\tau}+\bar{\tau}')}{(1-\rho)^2}\right)^{\frac{1}{1+\gamma}}\right\}\cdot \frac{\underline{\tau}}{\underline{\tau}+\bar{\tau}'}. \tag{E.5}$$

The above inequality is trivially true if $1 < \left(\frac{\omega\underline{\tau}}{(1-\rho)^2}\right)^{\frac{1}{1+\gamma}} < \left(\frac{\omega(\underline{\tau}+\bar{\tau}')}{(1-\rho)^2}\right)^{\frac{1}{1+\gamma}}$. Furthermore, if $\left(\frac{\omega\underline{\tau}}{(1-\rho)^2}\right)^{\frac{1}{1+\gamma}} < \left(\frac{\omega(\underline{\tau}+\bar{\tau}')}{(1-\rho)^2}\right)^{\frac{1}{1+\gamma}} < 1$, then note that both sides of the equation simplify to:

$$\begin{aligned} \left(\frac{\omega\underline{\tau}}{(1-\rho)^2}\right)^{\frac{1}{1+\gamma}} &> \left(\frac{\omega(\underline{\tau}+\bar{\tau}')}{(1-\rho)^2}\right)^{\frac{1}{1+\gamma}}\cdot\frac{\underline{\tau}}{\underline{\tau}+\bar{\tau}'} \\ \iff \left(\frac{\underline{\tau}}{\underline{\tau}+\bar{\tau}'}\right)^{\frac{1}{1+\gamma}} &> \frac{\underline{\tau}}{\underline{\tau}+\bar{\tau}'}, \end{aligned} \tag{E.6}$$

which is clearly true for $\gamma \geq 0$.

Thus, it suffices to show the inequality for the case $\left(\frac{\omega\underline{\tau}}{(1-\rho)^2}\right)^{\frac{1}{1+\gamma}} < 1 < \left(\frac{\omega(\underline{\tau}+\bar{\tau}')}{(1-\rho)^2}\right)^{\frac{1}{1+\gamma}}$, where the expression simplifies to showing:

$$\left(\frac{\omega\underline{\tau}}{(1-\rho)^2}\right)^{\frac{1}{1+\gamma}} > \frac{\underline{\tau}}{\underline{\tau}+\bar{\tau}'}. \tag{E.7}$$

This is clearly true, as:

$$\left(\frac{\omega\underline{\tau}}{(1-\rho)^2}\right)^{\frac{1}{1+\gamma}} = \left(\frac{\omega(\underline{\tau}+\bar{\tau}')}{(1-\rho)^2}\right)^{\frac{1}{1+\gamma}}\cdot\left(\frac{\underline{\tau}}{\underline{\tau}+\bar{\tau}'}\right)^{\frac{1}{1+\gamma}} > \left(\frac{\underline{\tau}}{\underline{\tau}+\bar{\tau}'}\right)^{\frac{1}{1+\gamma}} > \frac{\underline{\tau}}{\underline{\tau}+\bar{\tau}'}. \tag{E.8}$$

□

Figure A.12: Model Prediction for Implied Persistence in Additional Treatment Conditions

This figure shows the theoretical prediction of the implied persistence for our experimental interventions to change what's on top of mind. We use $\underline{\tau}^0 = \underline{\tau}/\alpha$ and $\alpha = 0.6$. The black dotted line shows the model's prediction for implied persistence in the baseline experiment. The red solid line shows the prediction for the additional experiments described above.