Memory and Overreaction in Expectations*

Hassan Afrouzi
Columbia

Spencer Kwon
Harvard

Augustin Landier
HEC Paris

Yueran Ma
Chicago Booth

David Thesmar
MIT Sloan

September 21, 2020

Abstract

We study how biases in expectations vary across different settings using a large-scale randomized experiment, in which participants forecast stable random processes. The experiment allows us to control the data generating process and the participants' relevant information sets, so we can cleanly measure forecast biases. We find that forecasts display significant overreaction to the most recent observation. Moreover, overreaction is especially pronounced for less persistent processes and longer forecast horizons. We also find that commonly-used expectations models do not easily account for the variation in overreaction across settings. We provide a theory of expectations formation with imperfect utilization of past information. Our model closely fits the empirical findings.

*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, Andrei Shleifer and Mike Woodford, as well as conference participants at AFA and seminar participants at Babson College, BYU, Columbia, Duke, the Fed, HEC Paris, Yale, MIT, NBER, NYU, Stanford SITE, and University of Amsterdam. David Norris provided very skillful research assistance on this project. 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

Expectation formation plays a critical role in economics. The benchmark model is rational expectations, which assumes that agents process information optimally and without bias. Empirically, however, a vibrant stream of recent studies uses survey data to document systematic biases in expectations, with evidence of overreaction in some settings[1] and underreaction in others.[2] In particular, biases in expectations seem to vary across different settings, but evidence and theory about how and why are still relatively sparse. Knowledge about such variations is an important step towards a unified understanding of findings on expectation biases. In this paper, we offer new experimental evidence and a new theory on how expectation biases vary with the features of the data generating process (DGP) as well as the forecast horizon.

We begin with a large-scale randomized experiment to cleanly document the relationship between biases in expectations and features of the process. Our experimental approach allows us to address three major concerns in analyzing expectations using survey data. First, we can control the relevant information set of forecasters, which is not observable to the econometrician in survey data.[3] Second, we can control and vary the DGP, whereas it is very difficult for the econometrician to know or control the DGP in survey data. Finally, we can also control the payoff of forecasters, while in field data there can be concerns that forecasters have considerations other than forecast accuracy. Overall, the experiment helps us measure biases in forecasts precisely, trace out the structure of

[1] A large share of this research follows up on the insight of Shiller (1981) that asset prices move more than fundamentals. 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), and Barrero (2020) document extrapolation and overreaction in expectations of corporate earnings and stock returns; Bordalo, Gennaioli and Shleifer (2018) show over-optimistic forecasts of future credit spreads during credit market booms.

[2] These papers have roots in both macroeconomics and finance. Mankiw and Reis (2002), Coibion and Gorodnichenko (2012, 2015) present evidence of informational rigidity in inflation expectations; Abarbanell and Bernard (1992), Bouchaud, Krueger, Landier and Thesmar (2019), and Ma, Ropele, Sraer and Thesmar (2020) find underreaction in near-term earnings forecasts.

[3] One workaround is to predict forecast errors using forecast revisions, since revisions are supposed to be within the forecaster's information set (Bordalo, Gennaioli, Ma and Shleifer, 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.

these biases and variations across settings, and then investigate whether commonly-used models account for the key findings in the data.

In our experiment, participants make forecasts of simple AR(1) processes. They are randomly assigned to a condition with a given persistence level, drawn from the set $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$. Participants observe 40 past realizations of the process at the beginning, and then make forecasts for another 40 rounds. In each round, participants observe a new realization from the process, and report one- and two-period-ahead forecasts before the next round begins. In follow-up experiments, we also extend the forecast horizon and elicit five-period-ahead forecasts.

Our main empirical results are as follows. First, even though the process is simple and stable, rational expectations are strongly rejected in our data, consistent with previous research. In particular, forecasts in the data display strong overreaction to recent observations: they are systematically too high when the past realization is high, and vice versa. This pattern 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 and importantly, we find that forecasts feature 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 can calculate the persistence implied by participants' forecasts, and compare it to the actual persistence of the process. In our setting, this implied persistence is a clear measure of overreaction. We find that the implied persistence is close to one when the process is a random walk, and decreases when the actual process is more transitory, but only reaches 0.4 for i.i.d. processes (where the actual persistence is zero).

Third, we find that commonly-used expectations models in the literature do not perform very well in accounting for how biases vary with the type of process. For example, the older adaptive or extrapolative models do not generate enough variation in the forecast-implied persistence based on the actual persistence of the process. In contrast, more recent models such as constant gain learning (Evans and Honkapohja, 2001; Nagel and Xu, 2019) and diagnostic expectations (Bordalo, Gennaioli and Shleifer, 2018) adapt too much: they overreact too little for transitory processes. Diagnostic expectations, for instance, are the same as rational expectations for i.i.d. processes, which is not the case in the data.

In light of the failure of these models to account for the key empirical features, in particular the variation of overreaction across different settings, we provide a new framework for biases in expectations. We consider the problem of an agent who forms estimates of the long-run mean of the process. For each round of forecasting, she initially observes a context, such as the most recent realization of a process, which automatically forms the initial prior. Then, the agent decides how much additional past information to utilize, subject to a cost of retrieval. The set of information retrieved, which we call working memory, captures what is "on top of the mind" when agents make decisions. In our model, like in the experiment, forecasts tend to overreact, since the agent partially relies on the most recent observation to estimate the long-run mean of the process. This mechanism naturally creates more overreaction for transitory processes (e.g., in an i.i.d. process the most recent observation brings no information about the long-run mean). This direct effect is, however, partially counteracted by the costly retrieval of past information. As a result, in our model, as in the data, the forecast adapts partially to the properties of the true process, but the adaptation is imperfect: there is a stronger tendency to respond too much to recent realizations when the true process is less persistent. When we fit our model to the forecast data, it matches the key empirical patterns very closely, unlike other models used in the literature.

Finally, recent research also indicates that overreaction appears to be stronger when the forecast horizon is longer (see Bouchaud et al. (2019) and Bordalo et al. (2019) for evidence from analyst earnings forecasts, as well as Brooks, Katz and Lustig (2018), Wang (2019), and d'Arienzo (2020) for evidence from interest rate forecasts). We also document this pattern in our experimental data. Moreover, our model naturally generates more overreaction at longer horizons. The intuition is that longer horizon forecasts are more sensitive to the estimate of the long-run mean, so they are more affected if the estimate of the long-run mean responds too much to recent observations. We take this model prediction to the data, and find the model performs well along this dimension too. In particular, we use the model parameters estimated using one-period-ahead forecasts, and compute the model-based forecasts for longer horizons as non-targeted moments. For two-period-ahead and five-period-ahead forecasts that we have in the experimental data, the model lines up very closely with the empirical evidence.

Literature Review. Our work is related to three branches of literature. First, our empirical findings complement recent evidence from survey data discussed in the first paragraph. As mentioned before, while analyses using survey data are very valuable, they face major obstacles given that researchers do not know forecasters' information sets, payoff functions, and the DGP. A key contribution of our study is using a large-scale experiment to cleanly connect the properties of the process with the structure of expectation biases.

Second, our paper also contributes to the literature on experimental studies of forecasts (see for instance Assenza, Bao, Hommes and Massaro (2014) for a survey). Prior work in this area includes Hey (1994), Frydman and Nave (2016) and Beshears, Choi, Fuster, Laibson and Madrian (2013). 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 test models or analyze the term structure of forecasts. We offer an extensive review of the experimental literature in Table A.1. Overall, relative to existing research, we provide an experiment with a large scale, a wide range of settings, and diverse demographics; we also collect the term structure of forecasts. In addition, we use the experiment to test a number of commonly-used models and to provide a unifying picture of expectation biases across different settings, while prior studies tend to focus on testing a given type of model.

Finally, we contribute to the emerging literature which proposes portable and microfounded models of expectations formation that allow for deviations from rational expectations. The diagnostic expectations model of Bordalo, Gennaioli and Shleifer (2018) is a leading example, but it does not explain biases when the process is i.i.d. as mentioned above. 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 the perception component of belief formation (e.g., imperfect perception of recent observations), and there is perfect utilization of past information. Instead, our model emphasizes frictions in exploiting past information, which is key for generating overreaction.

Given the frictions in exploiting past information, our model is related to recent models on memory and belief formation. Bordalo, Gennaioli and Shleifer (2020b) and Bordalo, Coffman, Gennaioli, Schwerter and Shleifer (2020a) draw inspirations from representativeness (Kahneman and Tversky, 1972) and associative recall (Kahana, 2012). Enke, Schwerter and Zimmermann (2020) experimentally test the role of associative recall in belief formation. Wachter and Kahana (2020) present a retrieved-context theory for belief formation. Nagel and Xu (2019) and Neligh (2020) study applications of memory decay. The most closely related analysis is da Silveira, Sung and Woodford (2020): they present a model of noisy memory and show its predictions for the empirical findings in our experiment. In their model, past information is summarized by a memory state formed before each period, and imprecise memory leads the agent to optimally put more weight on the latest observation, which generates overreaction. In our model, the agent decides the amount of past information to exploit depending on the current context. The costly retrieval of past information can reflect memory constraints, or "availability biases" more generally.

The rest of the paper proceeds as follows. Section 2 shows stylized facts from survey forecast data, and discusses the limitations of field evidence on overreaction, which leads us to conduct a simple experiment. Section 3 describes the experiment. Section 4 presents our main result — that overreaction is stronger for less persistent processes — and shows that commonly-used models fail at fitting it. We lay out our alternative model in Section 5, and show that it fits the data well. Finally, we discuss in Section 6 the additional prediction that overreaction is more pronounced at longer horizons. We also discuss modeling assumptions and robustness checks. Section 7 concludes.

2 Motivating Facts

To motivate our study, we first describe some stylized facts from survey forecasts of macroeconomic variables and corporate earnings. We show some intriguing patterns that emerge from survey forecast data, and discuss the key limitations of using survey data to analyze the variation of expectation biases across settings.

2.1 Overreaction and Process Persistence: Evidence from the Field

A major challenge for analyzing expectations using field data like surveys 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, which should incorporate news they respond to and should be part of their information sets. When forecasters overreact to information, forecast revisions at the individual level would overshoot: 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:

$$\underbrace{x_{t+h} - F_{i,t}x_{t+h}}_{\text{Forecast Error}} = a + b\,\underbrace{(F_{i,t}x_{t+h} - F_{i,t-1}x_{t+h})}_{\text{Forecast Revision}} + v_{it}, \tag{2.1}$$

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 forecasters display overreaction, b is expected to 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 more negative for processes with lower persistence. They interpret this pattern as an indication that overreaction tends to be stronger when the true 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 when the actual series is less persistent (i.e., more overreaction).

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 total assets, and the frequency is quarterly. Results are very similar if we use an annual frequency, or using earnings forecasts instead of sales forecasts.[4] 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 process ρ_i. Figure I, Panel B, shows a binscatter plot of the average b_i in twenty bins of ρ_i. Here, the majority of firms exhibit underreaction (as previously documented by Bouchaud et al. (2019)), but the key fact remains: the coefficient b_i is more negative when the actual sales process of the firm is less persistent.

These motivating facts in the field data point to the importance of understanding how subjective beliefs vary with the setting, which would be important for making progress in unifying existing empirical results and for guiding models of expectations.

2.2 Challenges in Field Data

The results from the error-revision regressions in field data, however, can be difficult to interpret unequivocally, for several key reasons.

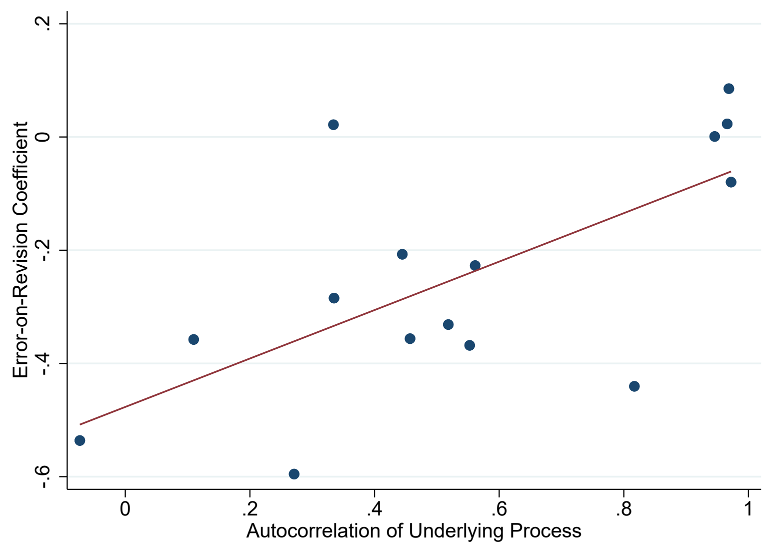
First, it is difficult to estimate b precisely for transitory processes when expectations are close to rational. In this case, revisions are close to zero, so the regression coefficient is not well estimated. As an illustration, in Figure A.1, Panel A, we show the error-revision coefficient b from simulations where we simulate forecasters under diagnostic expectations (Bordalo et al., 2018, 2020c) for AR(1) processes with different levels of persistence.

[4] 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).

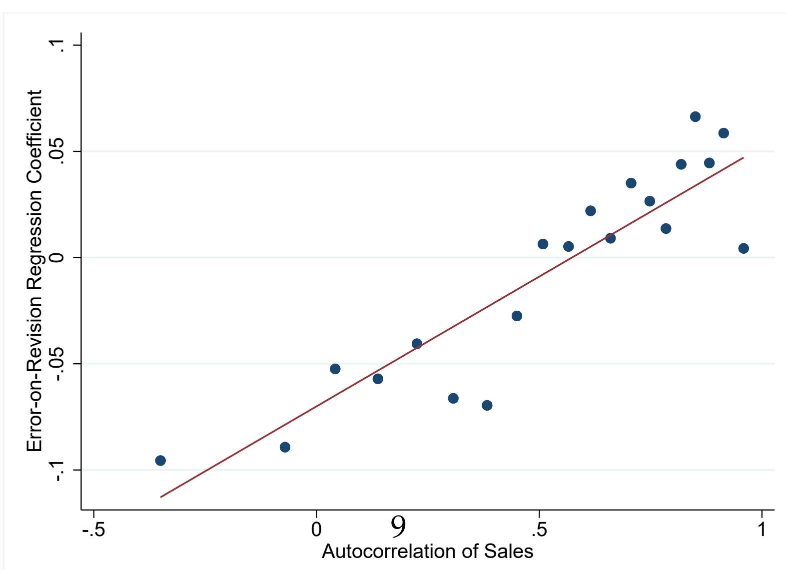
Figure I: Forecast Error on Forecast Revision Regression Coefficients

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 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 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



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 intervals become very wide when the process persistence is below 0.5.[5] 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 C.1).

Second, 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. 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. 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-hand side (forecast revision) and the left-hand side (forecast error) of the regression.

Taken together, the error-revision coefficient is a popular empirical measure in the field data, to circumvent issues arising from researchers not observing the forecasters' information sets and the DGP. It is inadequate, nonetheless, for measuring biases in expectations.

A more precise way to study the properties of subjective beliefs is to estimate the implied persistence from the forecasts ρ_h^s, which is the coefficient of regressing $F_t x_{t+h}$ on x_t when the process is AR(1). We can then compare it with the actual persistence ρ of the process. When $\rho_h^s > \rho^h$, 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). Figure A.1, Panel B, shows via simulations that this approach

[5]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 in period t, ρ is the AR(1) persistence, and ϵ_t is the shock to the process x_t 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.

is reliable for all levels of persistence. This alternative approach does not suffer from the shortcomings of the error-revision coefficient for two main reasons. First, the variance of the right-hand-side variable, the past realization, does not vanish to zero as ρ decreases. Second, the magnitude of ρ_h^s is much easier to interpret. For instance, ρ_h^s can be translated into a degree of overreaction by normalizing it using the rational sensitivity, ρ^h:

$$\zeta = \rho_h^s / \rho^h. \tag{2.2}$$

If $\zeta = 2$, then the subjective forecast responds twice as much as the rational forecast.[6]

Nonetheless, this approach is only meaningful if forecasters' information sets are restricted to past realizations of the process, and it requires that the DGP is truly AR(1). This is why we now turn to our experimental setting where we control both the forecasters' information set and the DGP.

3 Experiment Design

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

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

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." For MIT students, we send recruiting

[6] There is an approximate relationship between ζ and the error-revision coefficient. 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 ζ. Accordingly, a negative error-revision coefficient, often interpreted as evidence of overreaction, implies $\zeta > 1$, i.e., overreaction of the subjective belief to the latest observation.

emails to all students with a link to the experimental interface.

3.1 Experimental Conditions

There are three 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 6 values of ρ: $\{0, .2, .4, .6, .8, 1\}$. 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 270 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 $\rho \in \{.2, .4, .6, .8\}$. Experiment 3 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 providing 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 $\rho \in \{.2, .4\}$. 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.” At the end of the experiment, we further ask students questions testing their prior knowledge of AR(1) processes.[7]

[7] We do not disclose the values of μ and ρ, since the objective of our study is to understand how people

We focus on AR(1) processes because they are simple and therefore make the definition of rational expectations relatively clear. They are easy to learn as discussed more in Section 4. In addition, as Fuster, Laibson and Mendel (2010) point out, in finite samples, ARMA processes with longer lags are difficult to statistically tell apart from AR(1) processes. Finally, as discussed in Section 2.2, it is also straightforward to assess the degree of overreaction in this setting.

3.2 Payments

We provide fixed participation payments and 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 to earn a total score of about 2,800.[8] 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 and 2), 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.

Appendix Table A.3 shows participant demographics and other experimental statistics. Overall, MTurk participants are younger and more educated than the U.S. popula-

form forecasting rules; directly providing the values of μ and ρ would make this test redundant.

[8] $E(1 - |x_{t+1} - F_t|/\sigma)$ 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%.

tion. The mean duration of the experiment is about 15 minutes, and the hourly compensation is in the upper range of tasks on MTurk.

4 Main Empirical Findings

In this section, we present the main empirical findings from the experiment. In Section 4.1, we present the key stylized facts, connecting to the field data evidence discussed in Section 2. In Section 4.2, we then analyze whether commonly-used models of expectations are in line with these key facts.

4.1 Basic Fact: More Overreaction for More Transitory Processes

We begin by presenting the basic facts from our experiments. Figure II, Panel A, shows that the feature in SPF and IBES data discussed in Section 2 also holds in our experiment. Using data from Experiment 1, we have AR(1) processes with persistence from 0 to 1, and we run the error-revision regression in Equation (2.1), as we did on field data, for each level of persistence. As before, the y-axis shows the error-revision coefficient, and the x-axis shows the persistence of the process. Like in the field data, we see that the coefficient b is more negative for transitory processes.

Given the limitations of the error-revision regression approach explained in Section 2, a natural and more precise alternative in our experiment is the persistence implied by the forecast. The implied persistence is measured as the coefficient ρ_1^s in the regression:

$$F_{it}x_{t+1} = c + \rho_1^s x_t + u_{it}, \tag{4.1}$$

estimated in the panel of individual-level forecasts, for each level of AR(1) persistence ρ.[9] As the Full Information Rational Expectation (FIRE) is given by ρx_t, the difference between ρ_1^s and ρ provides a direct measure of the extent of overreaction. This measure is

[9] As in Bordalo et al. (2020c), we can also estimate the error-revision coefficient for each forecaster, and take the mean or median coefficient for each level of ρ. Similarly, we can estimate the implied persistence for each forecaster, $\rho_{1,i}^s$, and take the mean or median for each level of ρ. The results are very similar.

reliable for AR(1) processes as we show in Section 2, and forecasters' information sets are relatively clear in the experiment.

In Figure II, Panel B, we plot the implied persistence ρ_1^s against the true ρ. We see that when $\rho = 1$, ρ_1^s is roughly 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 roughly 0.45 (the sensitivity of the subjective forecast to x_t is much larger than the rational benchmark).[10]

Overall, in the experiment, by explicitly controlling for the DGP and the forecasters' information sets, we can establish clearly that overreaction is stronger for more transitory processes.

FIRE vs. In-Sample Least Square Learning. The comparisons above used the FIRE benchmark of true ρ. The results are very similar if we instead use in-sample least square learning as the rational benchmark. 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.

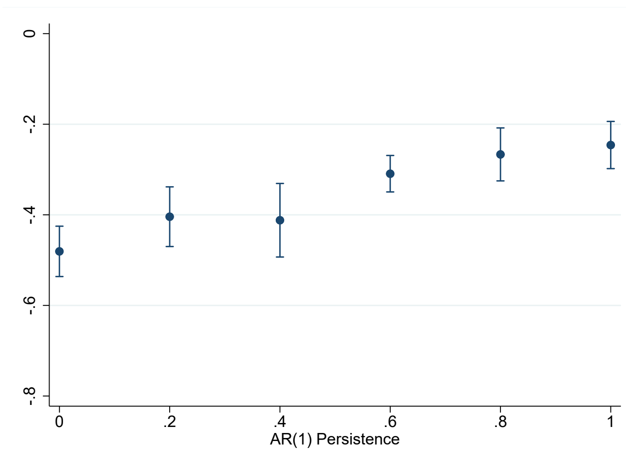
In our data, the difference between $\widehat{E}_t x_{t+h}$ and FIRE is small. The top panel of Appendix Figure A.4 shows that the mean squared difference between these two expectations is small, and does not decrease much after 40 periods. This is because our AR(1) processes are very simple, and a few 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

[10] We can also compute the ratio of relative overreaction $\zeta = \frac{\rho_h^s}{\rho^h}$ as defined in Equation (2.2). Internet Appendix Figure A.3 plots the value of ζ for each level of ρ (except when $\rho = 0$ where ζ 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.

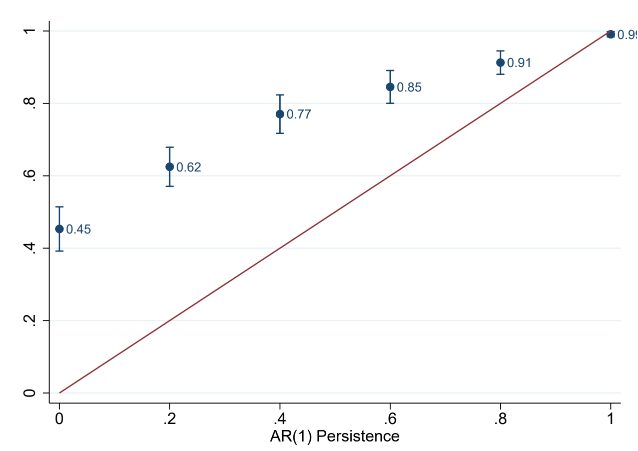
Figure II: Overreaction and Persistence of Underlying Process: Experimental Data

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^s from $F_{it}x_{t+1} = c + \rho^s x_t + u_{it}$ for each level of AR(1) persistence ρ. The y-axis plots the implied persistence ρ^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



least square learning is about the same as the true ρ. Accordingly, in the rest of the paper we use FIRE in our baseline definitions, but all the results are very similar if we use the in-sample least square $\widehat{E}_t x_{t+h}$ instead.

Effect of Linear Prior. We also analyze whether explicitly providing a linear prior 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). In Appendix Figure A.5, we show that whether this information was provided has no discernible impact 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. Appendix 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.

Stability across Demographics. Figure A.6 in the Appendix 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.

4.2 Testing Models of Expectations

We now use the data from our experiments and the key fact above to examine the performance of expectation formation models.

A. Models of Expectations

We begin by laying out commonly-used models of expectations below.

Backward-Looking Models

We begin with older "backward-looking" models, which specify fixed forecasting rules based on past data and do not incorporate properties of the process (i.e., 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. Since these models contain rational expectations, the term structure of expectations is more naturally defined.

3. Full information rational expectations

Full information rational expectations (FIRE) is the standard 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 experimental setup, where recent realizations are shown in real time, such frictions may correspond to noisy perception. These models typically have the following recursive definition:

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

where $E_t x_{t+h}$ is FIRE, and $\lambda \in [0, 1]$ depends on the noisiness of the signal. $\epsilon_{it,h}$ also comes from the noise in the signal.

Alternatively, this formulation could also represent anchoring on past forecasts. This formulation is used in Bouchaud et al. (2019) to model earnings 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 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 a version of LS learning where weights decrease for observations further in the past (Malmendier and Nagel, 2016). We use the 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; Nagel and Xu, 2019).

Other Models

The above list leaves out three classes of models in the literature: simple bounded rationality models, learning with nonlinear Bayesian priors, and natural expectations (Fuster, Laibson and Mendel, 2010; Fuster, Hebert and Laibson, 2012). The reason is we do not find evidence for these models in our data, by design or by outcome, as we explain below.

First, a possible model for our key fact is the one in Gabaix (2018). He describes a model where the forecaster faces a range of possible processes with varying degrees of persistence. To limit computational cost, the boundedly rational forecaster anchors the true persistence to a default level of persistence ρ^d: $\rho^s = m\rho_i + (1-m)\rho^d$. In such a setting, forecasters would tend to overreact to processes that are less persistent than average, and underreact to processes that are more persistent than average. This model has several limitations in our setting. First, it predicts underreaction for processes with high persistence, which we do not find in the data. Second, it is not clear how m and ρ^d are formed. Furthermore, models that solely work through misperceptions of the persistence parameter would predict diminishing overreaction for longer horizons, which is also not the case in the data as we discuss more in Section 6.

Second, we find no evidence of nonlinear priors in our data. Nonlinear priors may arise, for instance, because of beliefs in regime switches (Barberis, Shleifer and Vishny, 1998; Rabin, 2002; Rabin and Vayanos, 2010). 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. Overall, it does not appear that our empirical findings come from nonlinear priors.

Third, for natural expectations, the key observation is that forecasters may have difficulty differentiating processes with hump-shaped dynamics from simpler processes in finite samples (e.g., differentiating AR(2) or ARMA(p,q) from AR(1)), even based on statistical tools like BIC. In our tests, the emphasis is not the difficulty in detecting long-term mean-reverting processes in-sample. We focus instead on deviations from rational expectations for the simplest processes, an AR(1) which has much more simple dynamics.[11] Yet, even in this case, we find biases that have a clear structure.

B. Estimating Models of Expectations

We now estimate the six models described above on one-period ahead expectations data (i.e., with $h = 1$). We pool data from all conditions of Experiment 1 (i.e., with $\rho \in \{0, .2, .4, .6, .8, 1\}$). 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

[11] 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.

constant gain learning model is estimated by minimizing, over the decay parameter, the mean squared deviation between 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 strongly rejected, for at least two reasons. One is that FIRE has the lowest explanatory power of forecast data. The other is that 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%.

Second, most models point to strong signs of overreaction. The adaptive model features overreaction through the fact that the loading on the past realization x_t is very high (.83). This corresponds to overreaction whenever ρ is less than .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 past realization effect x_t, whose coefficient is estimated to be .93. The diagnostic expectations model has a θ of .34, which indicates strong overreaction (forecasts react 34% "too much" to the last innovation).[12] 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 statistically significant at $.14^{***}$, a magnitude consistent with earlier analyses on individual analyst EPS forecasts (Bouchaud et al., 2019). This finding suggests

[12]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).

that there is some anchoring on the level of past forecasts, in addition to overreaction to the recent realization.

C. Do Models Match the Relationships in the Data?

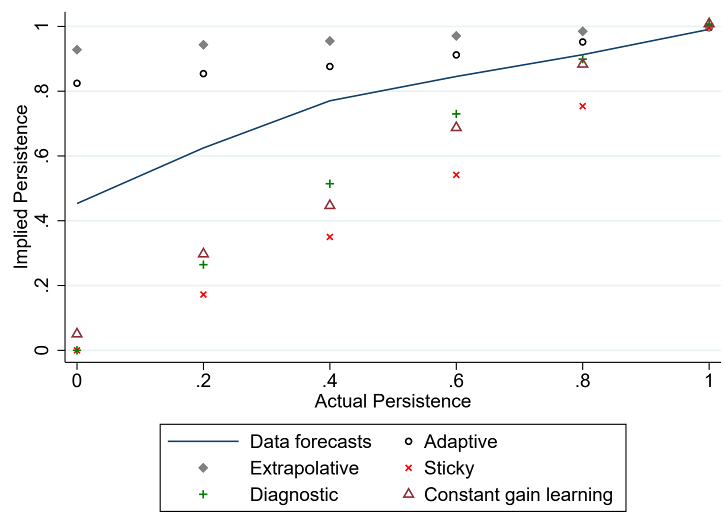
We first ask how the estimated models fit our key fact that overreaction is stronger for more transitory processes (our Figure II). We start with the pattern on the implied persistence, which is the most intuitive one. In Figure III, we compute the persistence implied by forecasts based on the five models estimated above. For each model m and for each observation in our data, we compute the predicted forecast $\widehat{F_t^m x_{t+1}}$, using the parameters in Table A.5. We then group observations per level of $\rho \in \{0, .2, .4, .6, .8, 1\}$. 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 III, the solid line represents the implied persistence based on actual forecasts (same as Figure II, Panel B). The dots represent the forecast-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 list of commonly-used models performs quite poorly for transitory processes. Backward-looking expectations models generate "too much" overreaction for transitory processes, while on the contrary, most forward-looking models do not generate enough overreaction. By definition, diagnostic and sticky expectations generate no overreaction for transitory processes (the forecast implied persistence according to these models is equal to zero). 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 forecasting data, do not seem to decrease fast enough.

To connect with results in field data and for completeness, we also report in Appendix Figure A.7 the error-revision coefficients based on the models. Again, the solid line represents experimental data (same as Figure A.7, Panel A) and the dots represent predictions from estimated models. In this figure 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 III.

Figure III: 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 $\rho \in \{0, .2, .4, .6, .8, 1\}$. 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.



For transitory processes, diagnostic and sticky expectations tend to lead to error-revision coefficients that are too high. Constant gain learning, on the contrary, generates a coefficient that is too negative.[13] Overall, the core message remains that commonly-used expectations models have trouble fitting the variation of expectation biases across settings with different levels of process persistence.

5 Model

Given the failure of commonly-used models to account for the empirical findings, we now introduce a model with a different approach, which provides a general framework for expectations formation that emphasizes recent data, context, and imperfect information utilization. We show that the model performs very well in matching the evidence

[13]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 III is immune to this problem.

described above.

5.1 Environment

Time is discrete and is indexed by $t \in \{0, 1, 2, \ldots\}$. There is an agent who tracks an exogenous stochastic process $\{x_t : t \geq 0\}$ and produces forecasts for the future realizations of this process at horizon h. The agent's payoff at any given time t depends on the accuracy of these forecasts and is given by:

$$-(F_t x_{t+h} - x_{t+h})^2, \tag{5.1}$$

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 $t + h$.[14]

We assume that x_t follows an AR(1) process with mean μ and persistence ρ:

$$x_t = (1 - \rho)\mu + \rho x_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2). \tag{5.2}$$

Working Memory. We assume that at the beginning of each period, the agent observes the context (defined as the most recent realization of x_t), decides whether to retrieve more data and form a working memory S_t, and uses the information in S_t to form her forecast about x_{t+h}.

Specifically, we assume that beliefs are formed based on a set of utilized data, which we refer to as working memory. The notion of working memory is a central component of our model. Although the agent may see many things, working memory in our model refers to the set of information that is "on top of the mind" when making decisions, consistent with the spirit of working memory in psychology research (Baddeley, 1983). Accordingly, a key feature of our model is it draws a distinction between data that

[14] It is important to 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 their payoff will be determined by the realization of the process at $t + h$. This is similar to the score function in the experiment with one slight difference that in the experiment, as discussed in Section 3.2, the score function does not have an exact quadratic form, to ensure that payments in the experiment are always non-negative. We use this standard quadratic form for simplicity of modeling, so we can derive closed-form solutions.

is potentially available, and the data utilized for the forecast. In our model, only a subset of all available data might be on top of the mind, which shapes the forecast. In the following, we model the process of retrieval, i.e., what comes to mind.

Formally, we assume that at the beginning of each period, the agent observes the most recent realization x_t costlessly, so that x_t is always in S_t. Furthermore, the agent can decide to retrieve more information from the history of past observations, but at a cost. If the cost is zero, then the model collapses to FIRE where all available data is retrieved as more data always leads to better forecasts. While this decision is trivial when retrieval is costless, in our setting, costly retrieval leads to a trade-off for information utilization. We assume that the cost of retrieved information is increasing and convex in *bits* of information retrieved by the agent. Formally, the cost of retrieval associated with S_t at time t, denoted by $C_t(S_t)$, is given by:

$$C_t(S_t) \equiv \omega \frac{\exp\left(2\ln(2) \cdot \gamma \cdot \mathbb{I}(S_t, \mu | x_t)\right) - 1}{\gamma}, \tag{5.3}$$

where $\omega \geq 0$ governs the overall cost of retrieval by shifting the function, and $\gamma \geq 0$ governs its convexity in Shannon's mutual information function ($\mathbb{I}(S_t, \mu | x_t)$) which measures the amount of information retrieved by the agent in units of bits after observing x_t.

The reason for assuming this functional form is that it embeds two useful special cases. First, it converges to be linear in $\mathbb{I}(S_t, \mu | x_t)$ when $\gamma \to 0$, which is the classic case that Sims (2003) assumed in introducing rational inattention and is widely used in that literature. Second, with a quadratic objective and a Gaussian posterior, it collapses to a quadratic cost in the precision of the agent's posterior when $\gamma = 2$, which is also used in the literature that assumes this precision is the choice variable of the agent (e.g., Angeletos and Pavan (2007)).[15]

Feasible Retrieval Set. Finally, to ensure that the agent cannot retrieve information beyond what is available at a given time, we assume that any retrieved information set should be independent of μ once we condition on the set of all available data at time t.

[15] For a formal derivation of these claims, see the proof of Lemma 1.

Formally, we define the set of feasible signals as follows.

Definition 1. Let $\bar{\mathcal{S}}_t$ be the set of all possible signals over μ at t. Then, given a history of available data at t, denoted by x^t, $s \in \bar{\mathcal{S}}_t$ is *feasible* to retrieve if it is independent of μ conditional on x^t. Formally, the feasible retrieval set for a given x^t is given by

$$\mathcal{S}_t(x^t) \equiv \{s \in \bar{\mathcal{S}}_t | \mathbb{I}(s, \mu | x^t) = 0\}. \tag{5.4}$$

Agent's Problem. Given the primitives of the problem at time t, the agent solves:

$$\begin{aligned} &\min_{S_t} \mathbb{E}\left[\min_{F_t x_{t+h}} \mathbb{E}\left[(F_t x_{t+h} - x_{t+h})^2 | S_t\right] + C_t(S_t)\right] \\ &s.t. \quad \underbrace{\{x_t\}}_{\text{observation}} \subseteq \underbrace{S_t}_{\text{working memory}} \subseteq \underbrace{\mathcal{S}_t(x^t).}_{\text{feasible retrieval set}} \end{aligned} \tag{5.5}$$

5.2 Characterization

We make two simplifying assumptions for our benchmark model. First, we assume that the agent's prior beliefs about the long-run mean μ after observing x_t is a normal distribution with mean x_t and precision $\underline{\tau}$.[16] Second, we assume that the agent knows the correct ρ for the process of x_t. As we discuss in Section 6, modeling frictions in beliefs about the long-run mean μ is the most parsimonious way to unify empirical evidence on expectation biases observed in the literature, while modeling frictions in beliefs about ρ does not seem sufficient.

Under these two assumptions, the problem simplifies to a simple choice of precision of the long-run mean estimate, summarized in the following Lemma:

Lemma 1. *For a set of available data $x^t \equiv \{x_\tau\}_{\tau=0}^t$, the agent's retrieval problem can be simplified*

[16] This can be obtained by assuming that the agent's prior before observing x_t is an improper uniform distribution.

to choosing the precision of the belief about μ:

$$\min_{\tau} \left\{ \frac{(1-\rho^h)^2}{\tau} + \omega \frac{\left(\frac{\tau}{\underline{\tau}}\right)^{\gamma} - 1}{\gamma} \right\} \tag{5.6}$$

$$s.t. \quad \underline{\tau} \leq \tau \leq \bar{\tau}_t \equiv var(\mu|x^t)^{-1}. \tag{5.7}$$

Proof. See Appendix C.2. □

The presence of $1-\rho^h$ in the objective function captures the fact that the agent is seeking to minimize prediction error over future outcomes x_{t+h}, not directly over the long-run mean μ. In the model, the assessment of the long-run mean is more important for longer horizons ($h \uparrow$), or processes with lower persistence ($\rho \downarrow$). The following proposition presents the solution to the retrieval problem.

Proposition 1. Suppose that the set of available data points is large enough that $var(\mu|x^t)$ is arbitrarily close to zero. Then the agent's optimal posterior precision about the long-run mean, $\tau^* = \text{var}(\mu|S^t)^{-1}$, is given by:

$$\tau^* = \underline{\tau} \max \left\{ 1, \left(\frac{(1-\rho^h)^2}{\omega \underline{\tau}} \right)^{\frac{1}{1+\gamma}} \right\}. \tag{5.8}$$

Moreover, the agent's forecast for x_{t+h} at time t, conditional on the true μ and realization of x_t, is distributed normally according to:

$$F_t x_{t+h}|(\mu, x_t) \sim \mathcal{N}(\mu_t, \sigma^2)$$

$$\mu_t \equiv \left(\rho^h + (1-\rho^h)\frac{\underline{\tau}}{\tau^*}\right) x_t \tag{5.9}$$

$$\sigma^2 \equiv (1-\rho^h)^2 \frac{1}{\tau^*}\left(1 - \frac{\underline{\tau}}{\tau^*}\right) \tag{5.10}$$

where we have normalized $\mu = 0$.

Proof. See Appendix C.3. □

5.3 Model predictions

We now explore the implications of our model for explaining the empirical evidence.

Overreaction. A key prediction of our model is that relative to rational expectations, forecasts under costly retrieval exhibit overreaction to the most recent observation. The reason is that the agent relies on the latest observation to predict the long-run mean of the process. This is a fundamental difference between our model and models of sticky information (which may use similar modeling techniques). In sticky information models, agents are fully aware of the past but some of them do not have access to the most recent observation, which can result in underreaction in the sense that forecasts rely more on the past than the present. In our model, agents are fully aware of the most recent observation and they have to decide whether to retrieve past data or not, which results in overreaction in the sense that forecasts rely more on the present than on the past data. To visualize this algebraically, we rewrite the equations of Proposition 1 as:

$$F_t x_{t+h} = \underbrace{E_t x_{t+h}}_{\text{rational forecast}} + \underbrace{(1-\rho^h)\min\left\{1, \left(\frac{\omega \underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}\right\} x_t}_{\text{overreaction}} + \underbrace{\varepsilon_t}_{\text{retrieval noise}} . \tag{5.11}$$

which shows that the bias relative to the rational benchmark has the sign of x_t, indicating systematic overreaction.

Comparative Statistics. In addition to predictions about overreaction in general, our model also predicts that the degree of overreaction varies with the persistence of the process. The reason is that for less persistent processes, the predictability of the long-run mean based on the most recent observation is lower and the agent needs to rely more on costly retrieval rather the most recent observation. The following proposition provides comparative statistics with respect to the parameters of the model.

Proposition 2. Consider the regression estimating the implied persistence ρ_h^s from the forecasts:

$$F_t x_{t+h} = c + \rho_h^s x_t + u_t, \tag{5.12}$$

and let $\Delta \equiv \rho_h^s - \rho^h$ denote the difference between asymptotic estimator of ρ_h^s in the data and the actual ρ^h of the process. Then,

1. $\Delta \geq 0$ with $\Delta = 0$ if and only if, either $\rho = 1$, or information retrieval is free ($\omega = 0$) 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 if the cost function is weakly convex in τ, which is true if and only if $\gamma \geq 1$.

Proof. See Appendix C.4. □

Furthermore, connecting this result to the measure of overreaction in Equation (2.2) yields the following corollary.

Corollary 1. *Consider the relative measure $\zeta \equiv \rho_h^s / \rho^h$. Then, $\zeta \geq 1$. Moreover, ζ is decreasing in ρ^h, for all values of ρ and h, if and only if $\gamma \geq 1$.*

Proof. See Appendix C.5. □

In summary, Proposition 2, along with its 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 implied and actual persistence, Δ, is positive (or equivalently, ζ, the relative measure of this gap, is greater than 1). The second result is that if the cost of retrieval is convex in the precision of the agent's forecast, the degree of overreaction, as measured by Δ or ζ, is larger for less persistent processes, as we observe in the data.

Moreover, the model provides two further testable predictions, which we discuss in more detail in Section 6. First, since what enters Δ or ζ is ρ^h, our results also imply that overreaction should be larger for longer-horizon forecasts (ρ^h is decreasing in h).

Second, ρ^h forms in a sense a sufficient statistic for overreaction: the implied persistence parameter should be similar in settings that share similar values of ρ^h.

Our model connects to recent work on memory and overreaction in belief formation. Wachter and Kahana (2020) construct a model of associative memory, which emphasizes the role of context in shaping retrieval. In that model, cued recall reinforces the association between two events, which may lead to overreaction. Nonetheless, the model does not address variation of overreaction process persistence and forecast horizon. da Silveira, Sung and Woodford (2020) present another approach of modeling overreaction through memory, which also assumes that memory is costly. In that model, agents decide what they want to remember in the future before an observation is revealed. In our model, the recent observation forms the key context, and agents decide to retrieve relevant information after an observation has been realized. In other words, while our model and the model in da Silveira, Sung and Woodford (2020) both deliver overreaction in posterior beliefs, the prior beliefs are anchored to different values: in our model, the priors are anchored to the present, namely the most recent observation; in da Silveira, Sung and Woodford (2020), in contrast, the priors are anchored to the past, which is given by the noisy memory state.

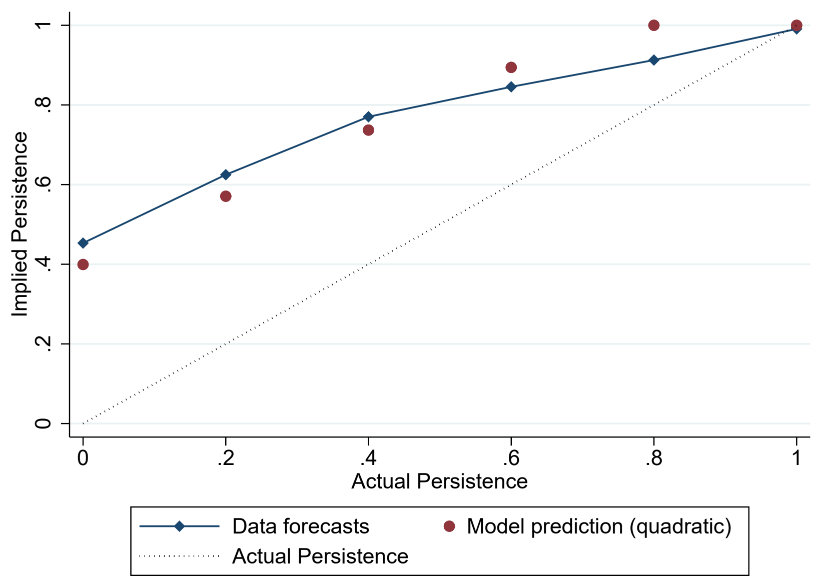
5.4 Model Fit

In the following, we present results on model fit for the case where the cost of retrieval is quadratic ($\gamma = 2$). We set $\gamma = 2$ in order to minimize the degrees of freedom in the model. We also present an alternative calibration in Section 6.2 where we jointly estimate γ with the other parameters of the model. We study the implied persistence in the data, and that predicted by our model when fitted to the realizations of x_t in the data. As before, the model is estimated by minimizing the mean-squared error (MSE) between the 1-period forecast predicted by the model for a given parameter (using the realizations of x_t in the data) and the 1-period forecast observed in the data.

Figure IV 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. The fit is much better compared to what we obtained in Figure III for the models in Section 4.2. Appendix Table A.6 further evaluates the model fit by calculating the MSE between ρ_h^s in by the model and ρ_h^s in the data, as well as the MSE between $F_t x_{t+h}$ in the model and $F_t x_{t+h}$ in the data. We calculate the MSE for our model and the models in Section 4.2. Confirming what is obvious visually, this MSE calculation also shows that our model has better performance than models discussed in Section 4.2.

Figure IV: Model Fit: Implied Persistence

This figure shows the forecast implied persistence ρ_1^s as a function of the objective persistence ρ. The implied persistence ρ_1^s is obtained by regressing $F_t x_{t+1}$ on x_t. The blue line represents the results in the forecast data. The solid red dot represents ρ_1^s from our model.



Finally, we discuss the intuition behind the better performance of our model. The alternative models in Section 4.2 can be categorized into two groups. For the first group, namely, adaptive expectations and traditional extrapolation, the models place a fixed weight on past observations that do not vary with the actual persistence ρ. Consequently, with a given parameter, these models generate implied persistence that adapts too little to the situation (the curve is too flat). For the second group, namely, diagnostic expectations and noisy information/sticky expectations, the models rely on rational expectations of the future forecasts. In particular, they converge to rational expectations when the true persistence is zero. The dependence on rational expectations and the adaptation turn out to be too strong in low persistence conditions (the implied persistence curve is too steep). In our framework, due to costly retrieval of past information, the forecaster conflates part of the transitory shock with changes in the long-run mean of the process. The agent adapts, but only partially because retrieval is costly. This partial adaptation is what makes our model fit the data better than the alternatives when $\rho = 0$: it overreacts less than backward-looking models, but more than the other non-RE forward-looking models.

6 Further discussion

In this section, we present additional non-targeted results from our model about how overreaction varies with the forecast horizon. We then show the robustness of our model formulations to different functional forms. We finally discuss the relevance and significance of several modeling assumptions.

6.1 Additional Implications for Forecast Horizons

Some recent research suggests that overreaction in survey data is also more pronounced for forecasts of longer horizon outcomes. 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. Wang (2019) and d'Arienzo (2020) use professional forecasters' predictions of interest rates, and show that the error-revision coefficient is negative and significant for long-term interest rates, but not for short-term interest rates. Earlier work by Giglio and Kelly (2018) using asset prices also points to “excess volatility" of long-term outcomes relative to short-term outcomes. Brooks, Katz and Lustig (2018) documents the same fact on the term structure of interest rates.

As noted above in Proposition 2 and Corollary 1, our model predicts that the degree of overreaction increases with $1 - \rho^h$, so it naturally delivers more overreaction for longer-horizon forecasts.

In the following, we present results for different forecast horizons in our data and our model. We begin with the empirical results in our forecast data. In addition to the one-period ahead forecast ($F_t x_{t+1}$) that we focus on in Section 4, from Experiment 2 we also have data on the two-period ahead forecast ($F_t x_{t+2}$), as well as the five-period ahead forecast ($F_t x_{t+5}$). We cannot construct the error-revision coefficient for these long-horizon forecasts, which will require information about ($F_t x_{t+3}$) and ($F_t x_{t+6}$) that is not available. Instead, we can study the implied persistence ρ_h^s associated with the long-term forecasts ($F_t x_{t+2}$ and $F_t x_{t+5}$) by regressing $F_t x_{t+h}$ on x_t. To aid visualization, we can also renormalize ρ_h^s to the implied per-period persistence as $\rho^s = (\rho_h^s)^{1/h}$. Internet Appendix Figure A.8 shows ρ^s for h =1, 2, and 5. Consistent with the model's prediction that $1 - \rho^h$ is a sufficient statistic for overreaction, we see that overall, the impact of increasing h (which leads to a smaller ρ^h) is similar to the impact of decreasing ρ, so that different setting with the similar values of $1 - \rho^h$ exhibit the same amount of bias.

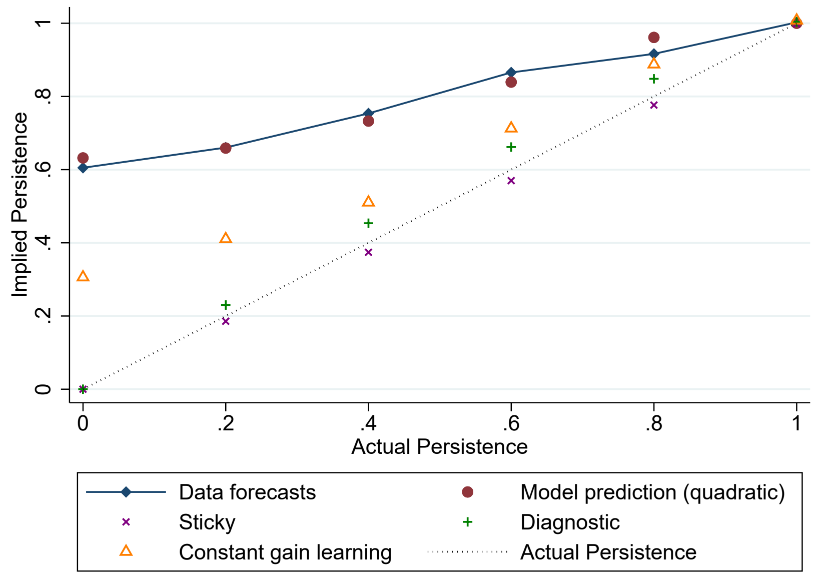
We can also ask how well our model fits the data for the longer horizon forecasts, which we show in Figure V, where Panel A studies the two-period-ahead forecast and Panel B studies the five-period-ahead forecast. Again, we show $\rho^s = (\rho_h^s)^{1/h}$ in the data, as well as the same quantity based on models discussed in Section 4.2 (dropping the adaptive model and the extrapolative model whose term structure of forecasts is not well defined) and based on our model. In particular, we fit all models using $h = 1$ (i.e., the model parameters are the same as those in Figure IV), so their performance for $h = 2$ and $h = 5$ are non-targeted. We see that the implied persistence according to standard models is too low: they do not produce sufficient overreaction for long horizon forecasts. Our model, on the other hand, performs quite well for the long-horizon forecasts, despite the moments being non-targeted. Appendix Table A.6 shows that our model also achieves the best fit in terms of MSE with respect to the forecasts in the data.

Overall, the data shows that overreaction is stronger for longer horizon forecasts. The commonly used models again do not seem to match the degree of overreaction for long horizon forecasts. Our model fits them quite closely.

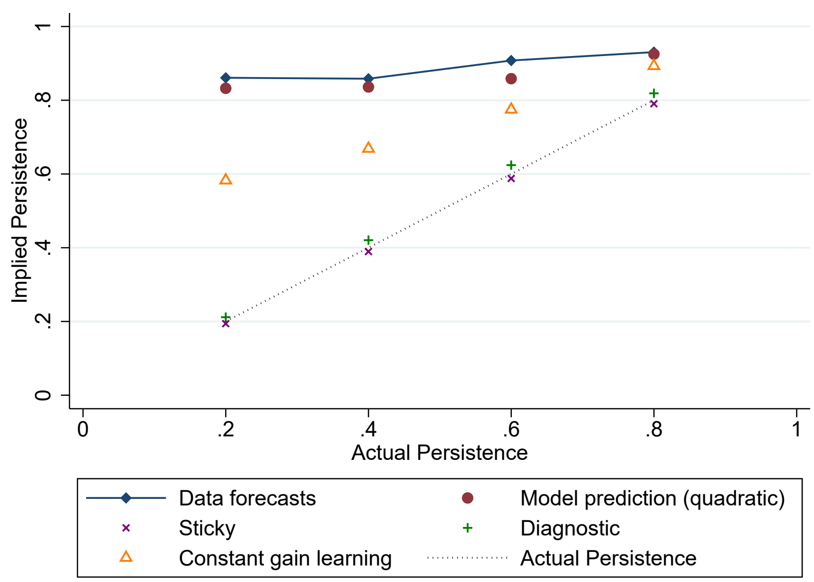
Figure V: Model Fit: Longer Horizon Forecasts

This figure shows the implied persistence ρ^s as a function of the objective persistence ρ. The subjective persistence ρ^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 and B show results for $h = 2$ and $h = 5$ respectively. The solid lines represent the value in the data. The solid red dot represents the value according to by our model. The dotted line is the 45-degree line.

Panel A. h=2



Panel B. h=5



6.2 Robustness of Model Formulations

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

A. Convexity and General Functional Form

We have assumed in our benchmark calibration that the cost of retrieval 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.9 in the Internet Appendix shows the fit of both exercises. 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 (with the optimal value of γ roughly equal to 10). Overall, however, we find that the model performance is not very sensitive to the exact value one picks for γ.

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

In our main model, we define $\underline{\tau}$ as the baseline precision the agent has regarding the long-run mean after seeing the most recent observation. For simplicity, we assumed $\underline{\tau}$ to be 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.10 in the Internet Appendix shows the fit of the alternative specification, and confirms that the model performs well in this case too.

C. Assumptions about ρ

In the model, we assume that the forecaster uses the correct ρ but may have biased estimates of the long-run mean μ. We make this modeling choice because biases about the mean are the most parsimonious way to account for the accumulating empirical evidence on predictable errors in forecasts. Biases about ρ (Gabaix, 2018; Angeletos, Huo and Sastry, 2020) may not be sufficient. For instance, such models do not necessarily account for the finding that overreaction is more pronounced in the long run than in the short run. In these models, the bias in ρ is attenuated for long-run forecasts as forecasters predict the long-run mean. On the other hand, our model, which focuses on inference about the long-run mean, does not have this problem.[17]

Overall, while we do not rule out that forecasters can directly use an incorrect ρ, we find that modeling biases about the mean μ is the most parsimonious way to capture biases in beliefs, and the variations with both the persistence of the true process and forecast horizons. This approach of modeling biases about the mean also has natural synergies with frictions of retrieving past information (if retrieval is costly then the mean can be estimated reasonably well). Thus our framework fits well with utilizing biases about the mean as a useful modeling setup.

D. Incentives

A possible question is whether one can test the effect of variation in incentives, or the relative trade-off between the cost of information retrieval and the benefit of obtaining accurate beliefs. While in principle one might ask whether these predictions can be tested in experiments, we have avoided doing so for several reasons. First, to obtain results that are statistically or economically strong, the magnitude of incentives may need to be substantially different across treatment arms, which can raise issues of fairness. For example, if an experiment randomly assigns participants to some conditions that pay ten or twenty times as much as other conditions, this design may be questionable to human

[17] Consider the example case of regressing the forecast error on the current realization. If the bias takes the form of using $\tilde{\rho}$ instead of ρ, then the coefficient of regressing forecast error of horizon h ($x_{t+h} - F_t x_{t+h}$) on the current realization x_t is $\tilde{\rho}^h - \rho^h$, which decreases with h. If the bias takes the form of using $\tilde{\mu}$ instead of the true mean, then the coefficient of regressing forecast error of horizon h on the current realization x_t is $(1-\rho^h)\beta_{\tilde{\mu}|x_t}$ (where $\beta_{\tilde{\mu}|x_t}$ is the regression coefficient of $\tilde{\mu}$ on x_t), which increases with h.

subject reviews and may antagonize potential participants when they read disclosures of payments in the consent form. Second, DellaVigna and Pope (2018) 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 retrieval. A large literature 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 suggest 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 the mind—are likely to be present.

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 the different sets of 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 in survey data.

We find that forecasts display significant overreaction: they respond too much to recent observations. Overreaction is particularly pronounced for less persistent processes and longer forecast horizons. We also find that commonly-used expectations models, estimated in our data, do not easily account for the variation in overreaction. Some predict too much overreaction when the process is transitory (e.g., adaptive expectations and simple extrapolation), while others predict too little (e.g., diagnostic expectations and constant gain learning).

We propose a new framework for understanding 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, but the utilization of past information can be costly and imperfect. As a result, forecasts adapt partially to the setting, but recent observations can have a disproportionate influence, resulting in overreaction. Over-adjusting the estimates of the long-run mean in response to recent observations also naturally 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 long-term forecasts that closely match what we observe in the data.

While our current model can provide a unifying framework for how the degree of overreaction varies with the setting, it does not generate underreaction and neither does our experimental evidence. Nonetheless, if there is noisy perception of the recent observation, then we can obtain underreaction too in the model. This is also a plausible reason for underreaction sometimes observed in survey forecast data (Coibion and Gorodnichenko, 2012; Bouchaud et al., 2019). Taken together, we hope that the theory and evidence in the paper contributes to the unification of 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.

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.

Angeletos, George-Marios and Alessandro Pavan, "Efficient Use of Information and Social Value of Information," *Econometrica*, 2007, *75* (4), 1103–1142.

_ , Zhen Huo, and Karthik A Sastry, "Imperfect Macroeconomic Expectations: Evidence and Theory," Working Paper 2020.

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 David, "Working Memory," *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 1983, *302* (1110), 311–324.

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 Implications of Managers' Beliefs," Working Paper 2020.

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

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.

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

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, Frederik Schwerter, and Andrei Shleifer, "Memory and Representativeness," *Psychological Review*, 2020, *Forthcoming.*

_ , 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.

Brooks, Jordan, Michael Katz, and Hanno Lustig, "Post-FOMC Announcement Drift in US Bond Markets," Working Paper 2018.

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

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.

da Silveira, Rava Azeredo, Yeji Sung, and Michael Woodford, "Optimally Imprecise Memory and Biased Forecasts," Working Paper 2020.

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

DellaVigna, Stefano and Devin Pope, "What Motivates Effort? Evidence and Expert Forecasts," *Review of Economic Studies*, 2018, *85* (2), 1029–1069.

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 2020.

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

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, Benjamin Hebert, and David Laibson, "Natural Expectations, Macroeconomic Dynamics, and Asset Pricing," *NBER Macroeconomics Annual*, 2012, *26* (1), 1–48.

— **, 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.

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.

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.

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.

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," Working Paper 2019.

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.

Shiller, Robert J, "Do Stock Prices Move Too Much to Be Justified by Subsequent Changes in Dividends?," *American Economic Review*, 1981, *71* (3), 421–436.

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

Thomas, M Cover and A Thomas Joy, "Elements of Information Theory," *New York: Wiley*, 1991, *3*, 37–38.

Wachter, Jessica A and Michael Jacob Kahana, "A Retrieved-Context Theory of Financial Decisions," Working Paper 2020.

Wang, Chen, "Under- and Over-Reaction in Yield Curve Expectations," Working Paper 2019.

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

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

This figure shows simulation results on the error-revision coefficient and the implied persistence coefficient. 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.

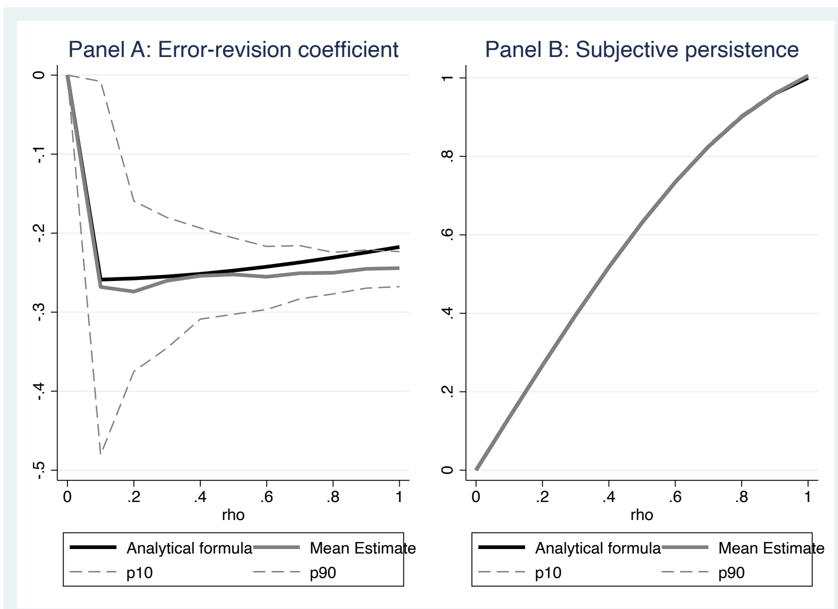


Figure A.2: Prediction Screen

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.

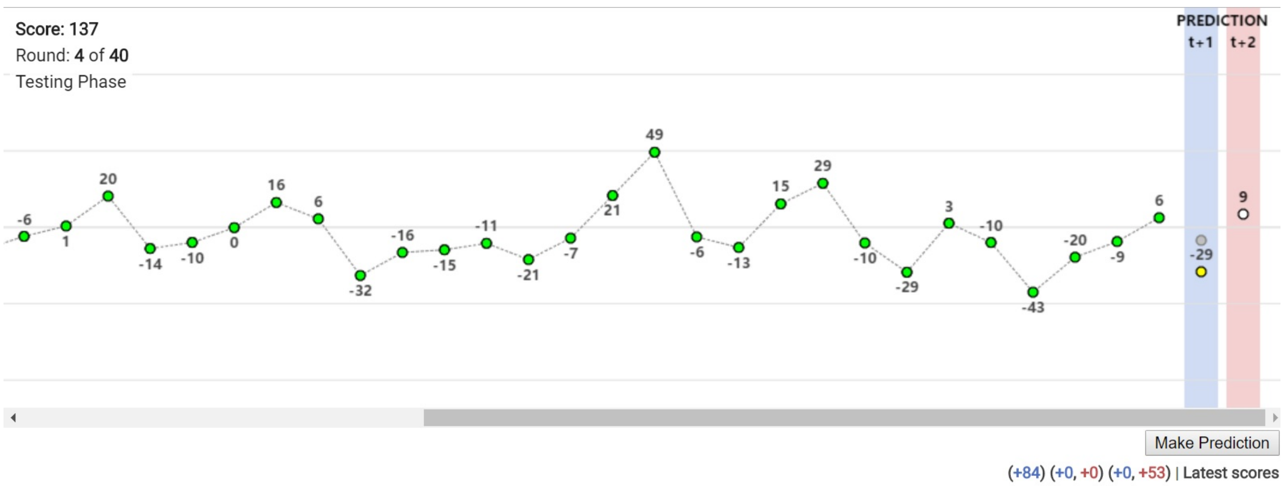


Figure A.3: Implied Persistence and Actual Persistence

We compute 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 relative to the actual persistence $\zeta = \rho_1^s / \rho$, i.e., the measure of overreaction, and the x-axis plots the AR(1) persistence ρ. The line at one is the FIRE benchmark.

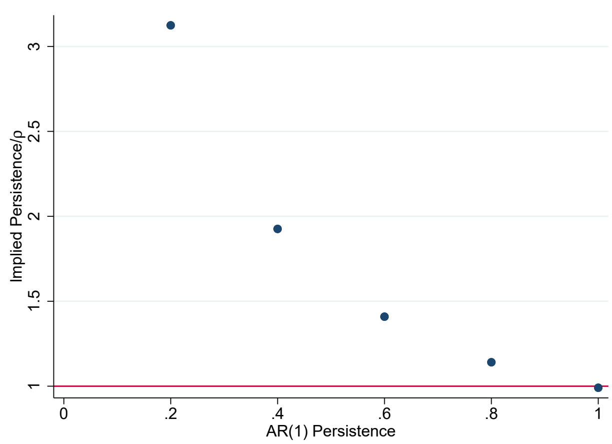
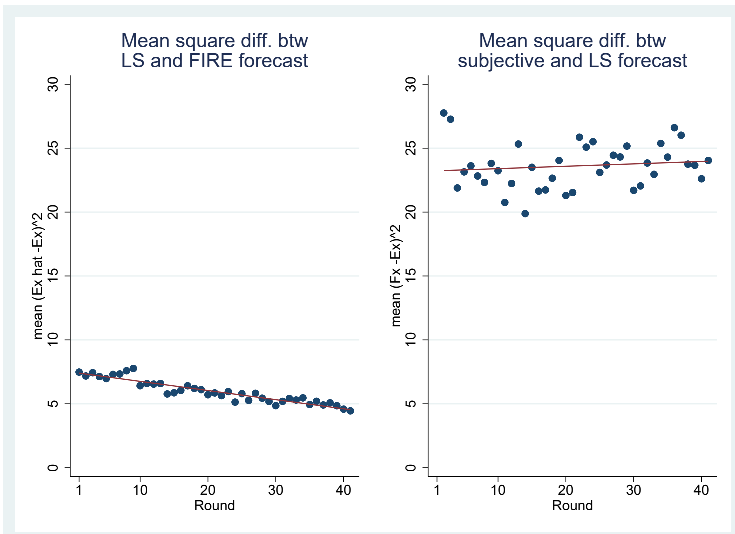


Figure A.4: Distance between Subjective Forecasts and Rational Expectations

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

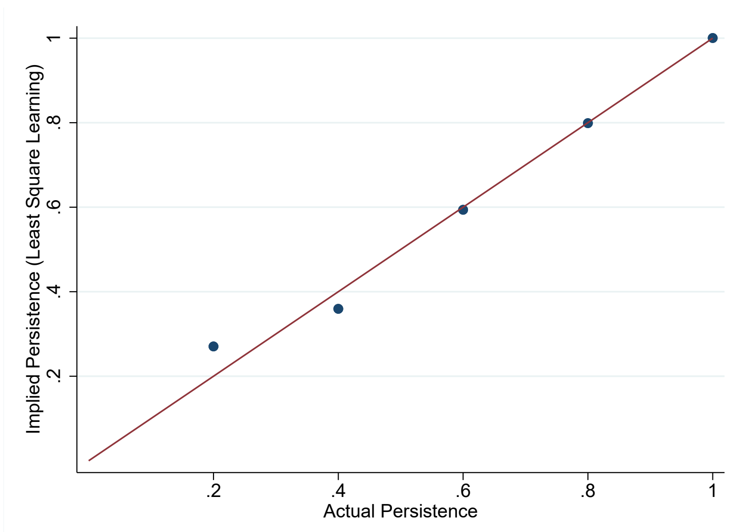
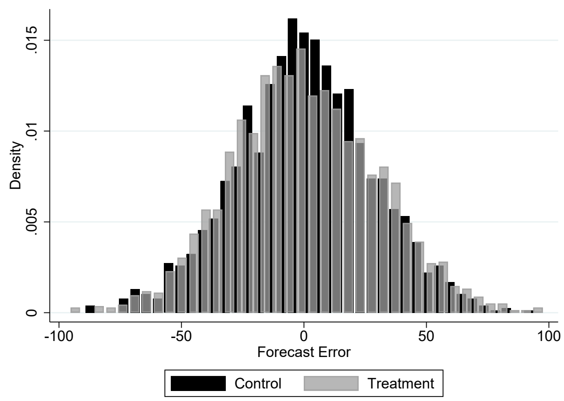


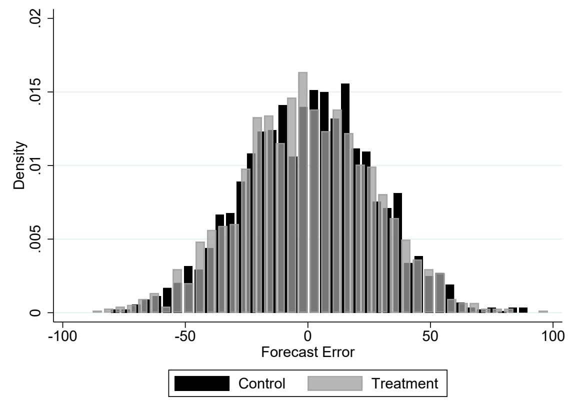
Figure A.5: Knowledge of Linear DGP and the Distribution of Forecasts

We use the data from Experiment 3 (MIT EECS), with 204 MIT undergraduates randomly assigned to AR(1) processes with $\rho = .2$ or $\rho = .6$. 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 ($x_{t+1} - F_t x_{t+1}$)

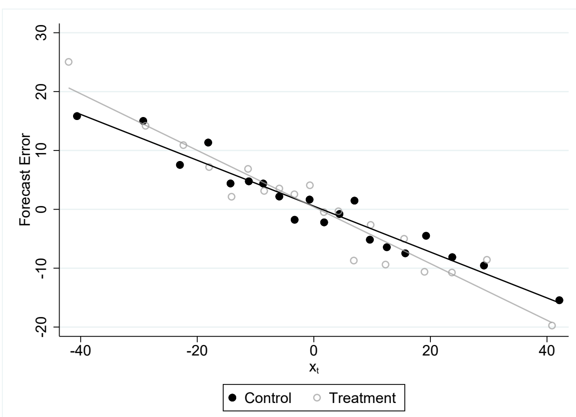


$\rho = 0.2$

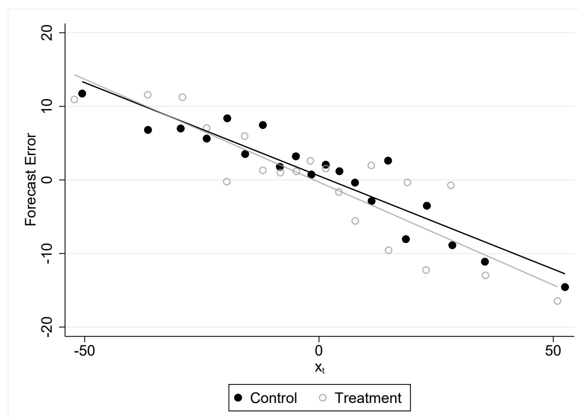


$\rho = 0.6$

Panel B. Forecast Error Conditional on x_t



$\rho = 0.2$

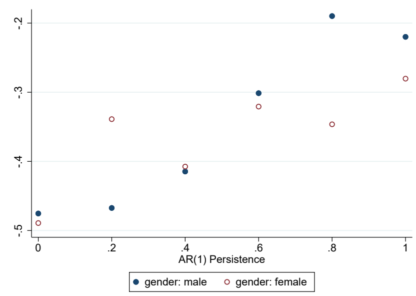


$\rho = 0.6$

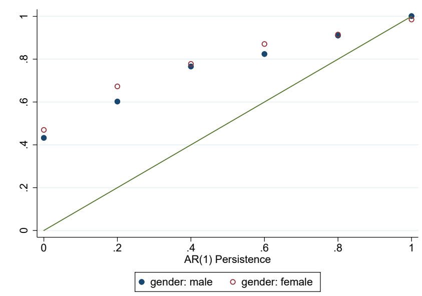
Figure A.6: Overreaction and Persistence of Process: Results by Demographics

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

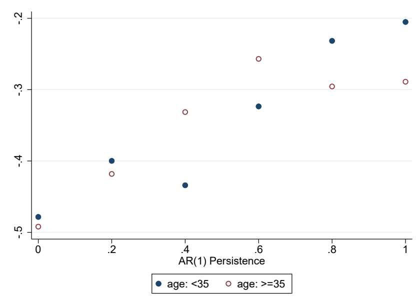


Error-revision coefficient by Gender

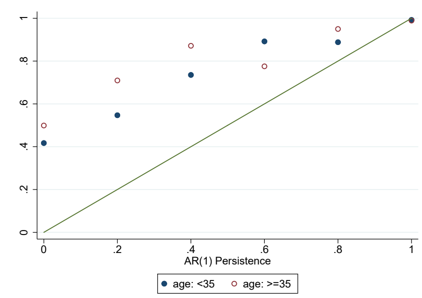


Implied Persistence by Gender

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

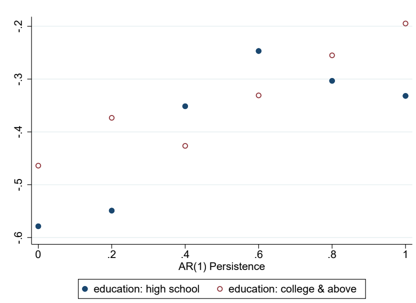


Error-revision coefficient by Age

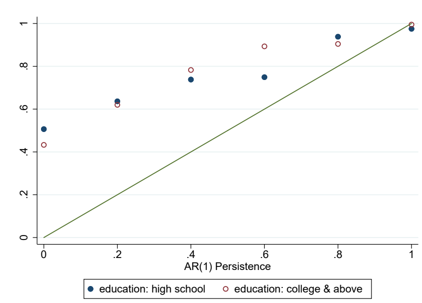


Implied Persistence by Age

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



Error-revision coefficient



Implied Persistence

Figure A.7: 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.

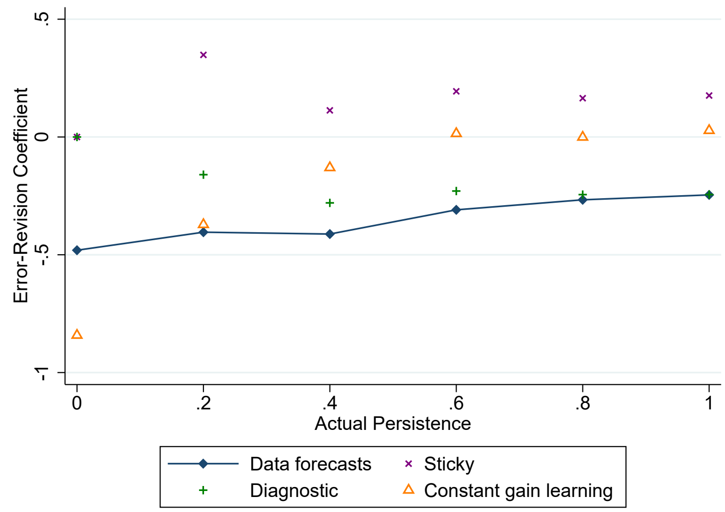


Figure A.8: Implied Persistence and Forecast Horizon

We report the forecast implied persistence for different horizons. For each horizon h and a given ρ, the x-axis is based on ρ^h, and the y-axis shows $\hat{\rho}_h$ which is the regression coefficient of $F_t x_{t+h}$ on x_t. Full dots correspond to $h = 1$ (from Experiment 1 where the one-period persistence $\rho \in \{0, .2, .4, .6, .8, 1\}$), which are identical to Figure II, Panel B. Empty circles correspond to $h = 2$ (also from Experiment 1). Crosses correspond to $h = 5$ and come from Experiment 3, where the one-period persistence $\rho \in \{.2, .4, .6, .8\}$. The solid orange line is the 45-degree line.

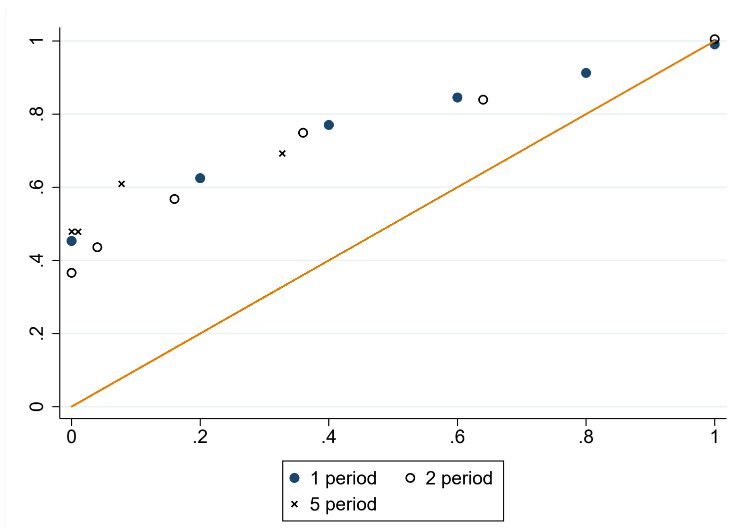
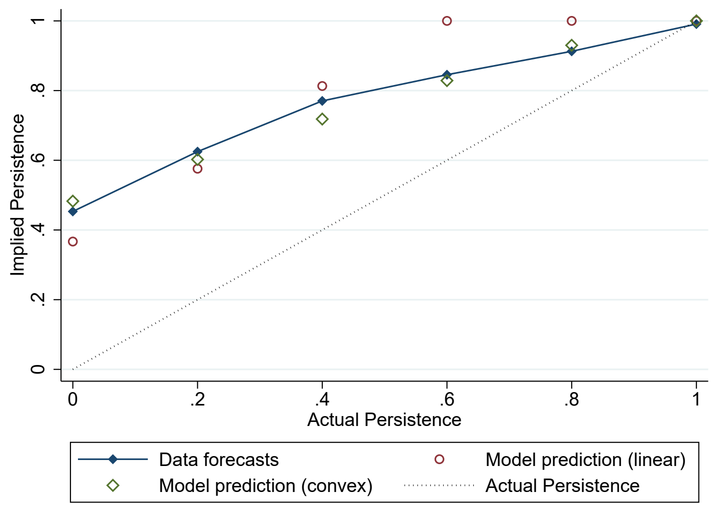


Figure A.9: Model Functional Form: Robustness Checks

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. $h = 5$

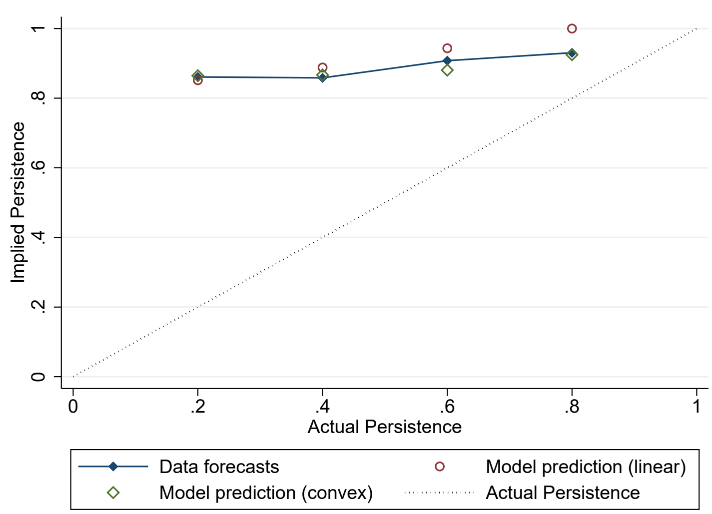
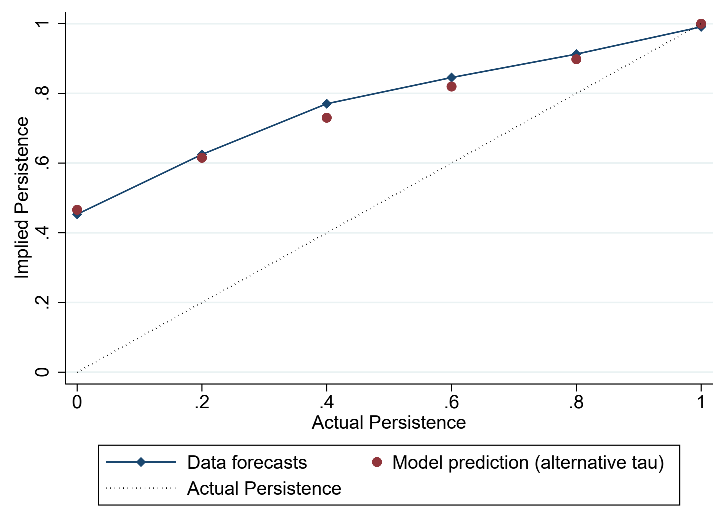


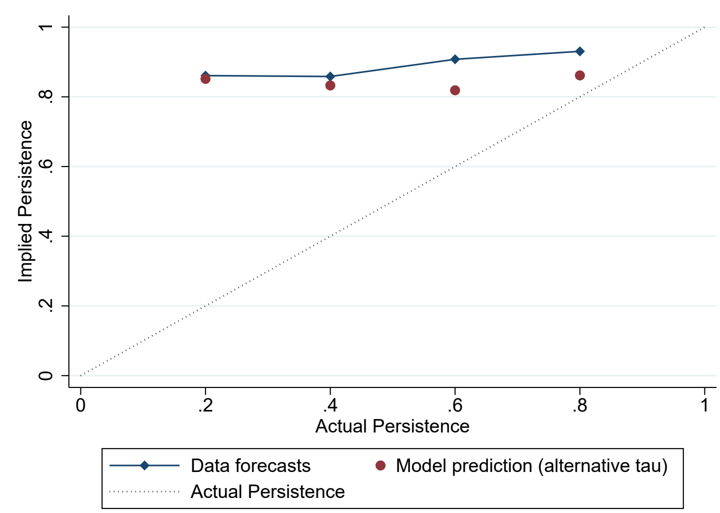
Figure A.10: Model Functional Form: Robustness Checks

This figure shows the model fit under alternative formulations of τ_I, 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. $h = 5$



B Appendix Tables

Table A.1: Experimental Literature on Expectations Formation

This table summarizes the experimental literature on expectations formation. 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) reports if all participants see the same draw or different draws of the same process. Column (7) is "Yes" if the data presented is presented as economic data or not. Column (8) shows that nearly all experiments feature some form of monetary incentives. Column (9) describes the format used to present the data (graphical or number list). Column (10) 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) Same draw	(7) Econ. background	(8) Monetary Incentives	(9) Format graph /text	(10) Forecast Horizon	(11) Model Tested
Schmalensee (1976)	23	26	27	$\rho \approx 1$	Yes	Yes	Yes	Both	1-5	Adaptive +Extrap.
Andreassen (1990)	77	5	5	$e^{\alpha t}$	Yes	Yes	No	Text	1	Extrap.
DeBondt (1993)	27	48	2	$\rho \approx 1$ non-rep.	Yes	Yes	Weak	Graph	7,13	Extrap.
Dwyer&al (1993)	70	30	40	$\rho = 1$	No	No	Yes	Both	1	Adaptive
Hey (1994)	50	20	48	$\rho \in \{.1, .5, .8, .9\}$	Yes	No	Yes	Both	1	Adaptive
Bloomfield &Hales(2002)	38	9	1	$\rho \approx 1$ non-rep	No	Yes	Yes	Both	1	Extrap.
Asparouhova et al(2009)	92	100	100	$\rho \approx 1$	Yes	No	Yes	Graph	1	BSV vs Rabin
Reimers &Harvey (2011)	2,434	50	Varies	$\rho \in \{0, 0.4, 0.8\}$	No	No	Yes	Graph	1	N/A
Beshears et al(2013)	98	100k	60	ARIMA (0,1,50)	No	No	Yes	Graph	1	Natural Expec.
Frydman &Nave(2016)	38	10	400	$\rho \approx 1$	No	Yes	Yes	Graph	1	Extrap.
This paper	1,600+	40	40	$\rho \in \{0, .2, .4, .6, .8, 1\}$	Both	Both	Yes	Both	1,2,5	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.

#	Short description	(1) persistence ρ	(2) AR(1) process constant μ	(3) volatility σ_ϵ	(4) Number of participants
Panel A: Experiment 1 – Baseline, MTurk					
A1	Baseline $\rho = 0$	0	0	20	32
A2	Baseline $\rho = 0.2$	0.2	0	20	32
A3	Baseline $\rho = 0.4$	0.4	0	20	36
A4	Baseline $\rho = 0.6$	0.6	0	20	39
A5	Baseline $\rho = 0.8$	0.8	0	20	28
A6	Baseline $\rho = 1$	1	0	20	40
Panel B: Experiment 2 – Long horizon, MTurk					
C1	Horizon: F1 + F5	0.2	0	20	41
C2	Horizon: F1 + F5	0.4	0	20	26
C3	Horizon: F1 + F5	0.6	0	20	31
C4	Horizon: F1 + F5	0.8	0	20	30
Panel C: Experiment 3 – DGP information, MIT EECS					
D1	Baseline	0.2	0	20	42
D2	Baseline	0.6	0	20	52
D3	Display DGP is AR(1)	0.2	0	20	70
D4	Display DGP is AR(1)	0.6	0	20	40

Table A.3: Summary Statistics

Panel A describes demographics of participants. Columns (1) and (2) provide information for participants in Experiment 1 (Baseline, MTurk); columns (3) and (4) for Experiment 2 (Long horizon, MTurk); columns (5) and (6) for Experiment 3 (Describe DGP, MIT EECS). 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

		(1)	(2)	(3)	(4)	(5)	(6)
		Experiment 1		Experiment 2		Experiment 3	
		Obs.	%	Obs.	%	Obs.	%
Gender	Male	151	55.9	201	60.9	88	43.1
	Female	119	44.1	129	39.1	116	56.9
Age	<= 25	36	13.3	44	13.3	197	96.6
	25-45	186	68.9	224	67.9	7	3.4
	45-65	44	16.3	57	17.3	0	0
	65+	4	1.5	5	1.5	0	0
Education	Grad school	26	9.6	42	12.7	0	0
	College	170	63.0	200	60.6	207	100.0
	High school	74	27.4	88	26.7	0	0
	Below/other	0	.0	0	.0	0	.0
Invest. exper.	Extensive	7	2.6	6	1.8	2	1
	Some	71	26.3	74	22.4	43	21.1
	Limited	100	37.0	129	39.1	138	67.7
	None	92	34.1	121	36.7	21	10.3
Taken stat class	Yes	110	40.7	144	43.6	-	-
	No	160	59.3	186	56.4	-	-

Panel B. Experimental Statistics

	Mean	p25	p50	p75	SD	N
Experiment 1						
Total forecast score	1,996	1,680	1,985	2,305	469	270
Bonus ($)	3.33	2.80	3.31	3.84	.78	270
Total time (min)	13.88	8.30	11.65	16.42	8.65	270
Forecast time (min)	7.10	4.49	5.77	7.85	4.39	270
Experiment 2						
Total forecast score	1,932	1,645	1,890	2,205	502	330
Bonus ($)	3.22	2.74	3.15	3.67	.84	330
Total time (min)	13.12	8.16	10.88	15.79	7.88	330
Forecast time (min)	6.78	4.59	5.75	7.74	4.05	330
Experiment 3						
Total forecast score	2,071	1,755	2,046	2,326	430	204
Bonus ($)	8.63	7.31	8.53	9.69	1.79	204
Total time (min)	18.47	7.57	10.02	14.09	37.67	204
Forecast time (min)	8.78	4.03	5.09	7.46	19.72	204

Table A.4: Effect of Knowing the Process

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)	Test of difference (p-value)
$\rho = .2$	0.56	0.65	0.14
$\rho = .6$	0.86	0.88	0.71

Table A.5: Estimations of Expectations 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 $\rho \in \{0, .2, .4, .6, .8, 1\}$). 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	$F_t x_{t+1} = \boldsymbol{\delta} F_{t-1} x_t + (1 - \boldsymbol{\delta}) x_t$	.17***	(.04)	.53
Extrapolative	$F_t x_{t+1} = (1 + \boldsymbol{\phi}) x_t - \boldsymbol{\phi} x_{t-1}$	-.07***	(.02)	.56
Panel B : Forward-looking models				
FIRE	$F_t x_{t+1} = E_t x_{t+1}$	-	-	.58
Sticky/noisy information	$F_t x_{t+1} = \boldsymbol{\lambda} F_{t-1} x_{t+1} + (1 - \boldsymbol{\lambda}) E_t x_{t+1}$	.14***	(.04)	.56
Diagnostic	$F_t x_{t+1} = E_t x_{t+1} + \boldsymbol{\theta}(E_t x_{t+1} - E_{t-1} x_{t+1})$	.34***	(.04)	.57
Constant gain learning	Rolling regression at t w/ weights: $w_s^t = \frac{1}{\boldsymbol{\kappa}^{t-s}}$	1.06***	(.01)	.56

Table A.6: Model Fit

This table shows the MSE between $\hat{\rho}^h$ 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	$h=1$		$h=2$		$h=5$	
MSE Type	$\hat{\rho}^h$	Forecast	$\hat{\rho}^h$	Forecast	$\hat{\rho}^h$	Forecast
	(1)	(2)	(3)	(4)	(5)	(6)
Current model	0.005	497.1	0.002	724.4	0.001	689.9
Adaptive	0.035	495.7	.	.	.	.
Extrapolative	0.064	527.3	.	.	.	.
Sticky	0.117	556.2	0.140	786.1	0.197	814.6
Diagnostic	0.069	521.2	0.115	758.0	0.177	803.3
Constant gain learning	0.067	526.8	0.039	749.5	0.033	736.3

C Proofs

C.1 Standard Errors of Error-Revision Coefficient

Proposition 3. 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 CLT, 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:

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

$\square$

Evidently, this ratio is bigger when the variance of x_i is smaller.
Under the error-revision approach, 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 approach 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.

The subjective persistence approach 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.

C.2 Lemma 1

Proof. The agent has two decisions: first, she decides what information to retrieve (choose $S_t \subseteq \mathcal{S}_t(x^t)$, and second, she chooses the optimal forecast $F_t x_{t+h}$ given the σ-algebra induced by S_t. We solve this backwards; namely, 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

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

It immediately follows that the firms' loss from an imprecise forecast is the variance of x_{t+h} conditional on S_t

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

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

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

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

where the second line follows from:

1. orthogonality of future innovations to S_t that follows from feasibility ($\varepsilon_{t+j} \perp \mathcal{S}(x^t), \forall j \geq 1$);
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 C.4 is independent of the choice for S_t. We can now rewrite the firms' problem as

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

$$s.t.\ \{x_t\} \subseteq S_t \subseteq \mathcal{S}(x^t), \tag{C.6}$$

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

The next step in the proof is to show that under the optimal information retrieval, the distribution of $\bar{x}|S_t$ is Gaussian. To prove this, we show that for any arbitrary $S_t \in \mathcal{S}(x^t)$, there exists another $\hat{S}_t \in \mathcal{S}(x^t)$ 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 $\mathcal{S}(x^t)$ and let $\hat{S}_t \supseteq \{x_t\}$ be such that

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

Such an $\hat{S}_t$ exists because $\mathcal{S}(x^t)$ is assumed to contain all possible signals on $\bar{x}_t$ 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,

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

A final observation yields us our desired result: by definition of the mutual information function in terms of entropy,[18]

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

Similarly,

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

It follows from these two observations that

$$C(\hat{S}_t) \le C(S_t) \Leftrightarrow \mathbb{E}[h(\bar{x}|\hat{S}_t)|x_t] \ge \mathbb{E}[h(\bar{x}|S_t)|x_t]. \tag{C.10}$$

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 holding only if both S_t and $\hat{S}_t$ are Guassian (see for example Cover Thomas and Thomas Joy (1991)).[19] This result implies that $C(\hat{S}_t) \le C(S_t)$. Therefore, for any arbitrary $S_t \subset \mathcal{S}_t(x^t)$ such that $\bar{x}|S_t$ is non-Gaussian, we have shown that there exists $\hat{S}_t \subset \mathcal{S}_t(x^t)$ that is (1) feasible and (2) strictly preferred to S_t and (3) $\bar{x}|\hat{S}_t$ is Gaussian.

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

$$\begin{aligned}
\mathbb{I}(\bar{x}; S_t|x_t) &= h(\bar{x}|x_t) - h(\bar{x}|S_t) && \text{(C.11)}\\
&= \frac{1}{2}\log_2(\text{var}(\bar{x}|x_t)) - \frac{1}{2}\log_2(\text{var}(x_t|S_t)). && \text{(C.12)}
\end{aligned}$$

For simplicity let us define $\tau(S_t) \equiv \text{var}(\bar{x}|S_t)^{-1}$ as the precision of belief about $\bar{x}$ generated by S_t and $\underline{\tau} \equiv \text{var}(\bar{x}|x_t)^{-1}$ as the precision of the prior belief of the agent about $\bar{x}$. It follows that

$$\begin{aligned}
\mathbb{I}(\bar{x}; S_t|x_t) &= \frac{1}{2\ln(2)}\ln\left(\frac{\tau(S_t)}{\underline{\tau}}\right), && \text{(C.13)}\\
C(S_t) &= \omega\frac{\exp(2\ln(2)\cdot\gamma\cdot\mathbb{I}(\bar{x}; S_t|x_t)) - 1}{\gamma} && \text{(C.14)}\\
&= \omega\frac{\left(\frac{\tau(S_t)}{\underline{\tau}}\right)^{\gamma} - 1}{\gamma}. && \text{(C.15)}
\end{aligned}$$

[18] 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: $h(Z) = -\mathbb{E}^Z[\log_2(f_Z(Z))]$.

[19] For completeness, here is a brief outline of 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:

$$\begin{aligned}
&\max_{\{f(x)\ge 0: x\in\mathbb{R}\}} -\int_{x\in\mathbb{R}} f(x)\log(f(x))dx && \text{(maximum entropy)}\\
s.t. &\int_{x\in\mathbb{R}} x^2 f(x)dx - (\int_{x\in\mathbb{R}} x f(x)dx)^2 dx = \sigma^2 && \text{(constraint on variance)}\\
&\int_{x\in\mathbb{R}} f(x)dx = 1. && \text{(constraint on } f \text{ being a PDF)}
\end{aligned}$$

Hence, the agent's problem can be rewritten as

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

$$s.t. \ \{x_t\} \subseteq S_t \subseteq \mathcal{S}(x^t). \tag{C.17}$$

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 at the time of information retrieval. Moreover, it has to be bounded above by $\text{var}(\bar{x}|x^t)^{-1}$ which 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. □

C.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 $\tau \leq \bar{\tau}(x^t) \equiv var(\mu|x^t)^{-1}$. 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

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

The next step is to find an optimal signal set $S_t \supseteq \{x_t\}$ that generates this posterior. 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

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

and

$$\sigma^2 \equiv var(\mathbb{E}[x_{t+h}|S_t]|\mu, x_t) = var(x_t|\mu, x_t) = 0; \tag{C.20}$$

2. if $\left(\frac{\omega\underline{\tau}}{(1-\rho^h)^2}\right) < 1$, then it means that the agent retrieves more information than what is revealed by the context x_t. Suppose a signal structure $\tilde{S}_t$ generates this posterior variance. By Lemma 1 this has to be Guassian. Our claim is that the set $\hat{S}_t \equiv \{x_t, \mathbb{E}[\mu|\tilde{S}_t]\}$ also generates this posterior. Note that elements of this set are also distributed according to a Gaussian distribution. To see the equivalence of the two sets, note that by the law of total variance,

$$\text{var}(\mu|x_t) = \text{var}(\mu|\tilde{S}_t) + \text{var}(\mathbb{E}[\mu|\tilde{S}_t]|x_t)$$
$$\text{var}(\mu|x_t) = \text{var}(\mu|\hat{S}_t) + \text{var}(\mathbb{E}[\mu|\hat{S}_t]|x_t),$$

but note that

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

Thus, it has to be that

$$\text{var}(\mu|\tilde{S}_t) = \text{var}(\mu|\hat{S}_t)$$

and the two sets generate the same posterior variance for the agent. Now, note that by Bayesian updating of Gaussians:

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

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

$$\text{cov}(\mu, \mathbb{E}[\mu|\tilde{S}_t]|x_t) = \text{var}(\mathbb{E}[\mu|\tilde{S}_t]|x_t) = \underline{\tau}^{-1} - \tau^{-1}, \tag{C.21}$$

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 $\tilde{S}_t$. 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

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

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

$$\begin{aligned}
&\mathbb{E}[\mathbb{E}[\mu|\tilde{S}_t]|\mu, x_t] = (1 - \frac{\underline{\tau}}{\tau})\mu + \frac{\underline{\tau}}{\tau}x_t \\
\Rightarrow &\mu_t \equiv \mathbb{E}[\mathbb{E}[x_{t+h}|\tilde{S}_t]|\mu, x_t] = (1-\rho^h)(1 - \frac{\underline{\tau}}{\tau})\mu + (1-\rho^h)\frac{\underline{\tau}}{\tau}x_t + \rho^h x_t.
\end{aligned} \tag{C.22}$$

Moreover,

$$\begin{aligned}
&\text{var}(\mathbb{E}[\mu|\tilde{S}_t]|x_t) = a^2\text{var}(\mu|x_t) + \text{var}(\varepsilon_t) \\
\Rightarrow &\text{var}(\varepsilon_t) = \frac{1}{\tau}(1 - \frac{\underline{\tau}}{\tau}) \\
\Rightarrow &\sigma^2 \equiv \text{var}(\mathbb{E}[x_{t+h}|\tilde{S}_t]|\mu, x_t) = (1-\rho^h)^2\text{var}(\varepsilon_t) = (1-\rho^h)^2\frac{1}{\tau}(1 - \frac{\underline{\tau}}{\tau}).
\end{aligned} \tag{C.23}$$

Plugging in the expression for τ from (C.18) into (C.22) and (C.23) and setting $\mu = 0$ gives us the expressions in the Proposition.

Combining Equations (C.19), (C.20), (C.18), (C.22), (C.23) and setting $\mu = 0$ gives us:

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

$$\sigma^2 = (1-\rho^h)^2\underline{\tau}^{-1}\max\left\{0, \left(\frac{\omega\underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}\left(1 - \left(\frac{\omega\underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}\right)\right\}. \tag{C.25}$$

□

C.4 Proposition 2

Proof. From Proposition 1 we can derive Δ as

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

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 $\mathrm{var}(\mu|x^t)$ is arbitrarily small. So $\Delta = 0$ if and only if either $\rho = 1$ or $\omega = 0$ and past information 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 increasig 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

$$1 - \frac{2}{1 + \gamma} \geq 0 \Leftrightarrow \gamma \geq 1. \tag{C.27}$$

□

C.5 Corollary 1

Proof. From Proposition 2 we have

$$\ln(\zeta) = \ln \left(1 + (\rho^{-h} - 1) \min \left\{ 1, \left(\frac{\omega \underline{\tau}}{(1 - \rho^h)^2} \right)^{\frac{1}{1+\gamma}} \right\} \right). \tag{C.28}$$

First of all, it is straightforward to see that the term in side the log in the rihgt-hand side is larger than 1, so perceived persistence is larger than actual persistence — in other words, ζ is a measure of overreaction. Moreover, for ζ to be decreasing in ρ^h it has to be the case that $(1 - \rho^h)^{1 - \frac{2}{1+\gamma}} / \rho^h$ is 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, it has to be the case that $\gamma \geq 1$. □