New Experimental Evidence on Expectations Formation [*]

Augustin Landier[†] Yueran Ma[‡]

David Thesmar[§]

December 7, 2018

Abstract

We measure belief formation in an experiment where participants are asked to provide forecasts of a stable and simple statistical process. We then estimate an empirical model that allows for under- and over-reaction, as well as stickiness. Our findings are threefold. First, the rational expectations hypothesis is strongly rejected in our setting, and we find little evidence of learning. Second, both extrapolation and stickiness patterns are statistically discernible in the data, but extrapolation quantitatively dominates. Third, our model coefficients are robust across different settings, including to changes in experimental setting: they do not depend on process parameters, individual characteristics, framing.

[*] Many thanks to Nick Barberis, Cary Frydman, Charles Nathanson, Andrei Shleifer and Michael Woodford, as well as seminar participants at Babson College, BYU, Columbia, Duke, the Fed, HEC Paris, Yale, MIT, NBER, 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."

[†] HEC Paris

[‡] Chicago Booth

[§] MIT Sloan and CEPR

1 Introduction

The way decision makers form expectations about future outcomes is at the very core of most economic models. Rational expectations assume that agents process information fully and in an unbiased way. Expectations formation in practice, however, may have imperfections: decision makers may over-react or under-react to information, leading to predictable forecast errors. A vibrant stream of recent studies aims to develop a better understanding of expectations formation using empirical evidence from survey data. These studies provide accumulating evidence of biases in expectations across many domains. A first branch of research emphasizes extrapolation and over-reaction: Agents tend to over-estimate the persistence of recent trends or recent shocks, which can lead to excessive movements in asset prices as reflected by the observations of Shiller (1981).[1] Another branch of research uncovers stickiness in expectations adjustment, as well as under-reaction to information.[2]

The rich set of evidence from survey data highlights that expectations formation in many economic settings appears imperfectly rational. However, survey data also face inevitable limitations. First, the underlying data generating process is not easy to pin down, which makes it challenging to precisely differentiate rational vs. irrational updating. Second, it is also challenging for econometricians to know forecasters' information sets, which adds to the difficulty of specifying the biases in updating. Third, there can be concerns of forecasters' strategic considerations.

In this paper, we analyze expectations formation in a simple, large-scale experiment where agents are incentivized to provide accurate forecasts of a random variable, drawn from a basic stationary statistical process (an AR(1) process). While we are aware of potential external validity concerns, the experiment has several advantages. First, we are able to specify the data generating

[1] Debondt and Thaler (1985), Amromin and Sharpe (2013), Greenwood and Shleifer (2014), Gennaioli et al. (2016), Nagel and Xu (2018), Bordalo et al. (2018a), and Barrero (2018) document such biases in expectations of corporate earnings and stock returns; Bordalo et al. (2018c) show over-optimistic forecasts of future credit spreads during credit market booms; Bordalo et al. (2018b) document over-reaction in professional forecasters' expectations of macroeconomic outcomes.

[2] Mankiw and Reis (2002), Coibion and Gorodnichenko (2012), and Coibion and Gorodnichenko (2015) present evidence of informational rigidity in inflation expectations; Abarbanell and Bernard (1992), Bouchaud et al. (2018), and Ma et al. (2018) find under-reaction in near-term earnings forecasts.

process. We inform agents that the data generating process is stable and endow them with enough observations to understand the features of the process. Second, the relevant information set for the forecasting task is simple and well-defined. Third, the task in the experiment is a pure forecasting exercise, where agents' incentives are clearly specified: they just need to make accurate forecasts of a simple process. Finally, the experiment also allows us to test the robustness of our findings in different settings. For instance, we implement the experiment among nearly two thousand participants from diverse demographic groups. Across different treatment conditions, we also vary parameters of the stochastic process, study forecasts at different horizons, and modify the framing of the process, among others. Taken together, the experiment provides us with a clean benchmark setting to examine robust features of expectations formation.

In the experiment, participants start by observing past 40 realizations of a stationary AR(1) process, and make forecasts of the future realizations of this process for 40 more rounds. In each round, they observe the actual realization and make new forecasts. In the baseline experiment, participants make forecasts of outcomes in $t+1$ and $t+2$ in each round; the persistence of the AR(1) process ranges from 0 to 1 across different treatment conditions, and participants are randomly assigned to one of the treatment conditions. In additional experimental conditions, we also study different forecast horizons and framing.

Our experiment generates a large panel of subjective forecasts and actual realizations of random processes across different treatment conditions. We use this panel to estimate an expectations formation model that allows for both extrapolation and stickiness, but nests rational expectations as a particular case. More precisely, by "extrapolation," we mean over-estimating the persistence of recent shocks; by "stickiness," we mean that current expectations may be overly influenced by previous expectations. The data from the experiment allows us to separately identify these two types of biases because we know the data generating process and we collect the term structure of expectations.

Our main findings are the following. First, the rational expectations hypothesis is strongly rejected in our data, consistent with previous empirical findings, even though the process is simple and stable. The rejection holds both for the majority of individual participants and for the average forecasts. In our main experiment, some 90% of the participants have a score below the one obtained by a rational learner, with forecasting scores on average 26% lower.[3] In addition, we find little evidence that subjective forecasts converge to rational ones after 40 rounds of testing.

Second, both extrapolation and stickiness are statistically discernible in the data. Specifically, expectations tend to be influenced by previous forecasts (stickiness), but also tend to exaggerate the impact of the most recent shock (extrapolation). Extrapolation dominates quantitatively, but stickiness propagates past biases.

Third, while expectations display consistent biases, they are not mechanical: they incorporate features of the true process, albeit imperfectly. Models that do not incorporate features of the true process and only use past information based on a deterministic rule do not perform well in fitting the data. This finding echos Bordalo et al. (2018c)'s observation of the "kernel of truth": expectations are biased but they do adapt to different contexts (and such biased expectations are not necessarily subject to the Lucas critique).

Fourth, our model explains average expectations very well (with an R^2 of 50%-60% depending on specifications). While individual forecasts have some heterogeneity and also contain some noise, we do not find that different agents use substantially different models in this simple setting.

Finally, our model coefficients are remarkably robust to the experimental setting. They are stable across different parameters of the process (the persistence and volatility of the AR(1) process), different ways of labeling the process (random process, GDP, inflation etc.), and across different demographic groups (participants from the general population across the US, as well as MIT undergrads in Electrical Engineering and Computer Science). The extent of stickiness is somewhat affected by the way we remind participants of their earlier forecasts, but in all cases, the amplitude of over-reaction is surprisingly stable.

Taken together, our results are broadly supportive of recent models of over-extrapolation such as the diagnostic expectations of Bordalo et al. (2018c). We find that over-extrapolating the

[3] We describe our methodology in more detail in Section 3.2. The "rational learner" here is an econometrician fitting a linear model on all available past data.

impact of recent shocks is significant and prevalent, although the most reliable functional form to capture the shock in the data is slightly different from the functional form in Bordalo et al. (2018c). Moreover, while the biases in Bordalo et al. (2018c) are transitory, we document stickiness in beliefs in our data, which can prolong past biases.

We also contribute to the experimental studies on expectations (see for instance Assenza et al. (2014) for a survey). Previous experimental work generally rejects simple forms of the rational expectations hypothesis. Using AR(1) processes, Hey (1994) rejects rational expectations and finds evidence that adaptive expectations have explanatory power for belief dynamics. Using binary outcomes, Frydman and Nave (2016) document extrapolative biases in both perceptual and economic decisions. Bloomfield and Hales (2002) and Asparouhova et al. (2009) find evidence of regime-shifting beliefs using pre-specified paths of outcomes or binary outcomes. As we discuss in more detail in Section 4, we do not find strong evidence of regime-shifting beliefs in our data, possibly because continuous outcomes lead to many more possibilities and streaks are less obvious. Beshears et al. (2013) study processes with short-run momentum and long-run mean reversion and find that participants generally under-estimate the long-run reversals. We offer a more exhaustive literature review in Table 1. A core contribution of our paper is the scale of the experiment. We have more than 1,600 participants relative to less than 100 in most existing studies. This allows us to explore a richer set of processes with different levels of persistence and a richer set of forecast term structure, both of which are important for differentiating biases. We also demonstrate the coexistence of extrapolation and stickiness.

In the following, Section 2 describes our experimental design. Section 3 describes the empirical model we use. Section 4 presents the results. Section 6 concludes.

2 Experiment Design

In the experiment, participants first read a consent form (shown in the Survey Appendix), with a brief description of the experiment, the payments, and the duration. Once participants agree to the consent form, they read instructions and start the experiment. In all tests, we first present participants with 40 historical realizations of a statistical process. Participants then forecast future realizations for 40 rounds. *After* the prediction task, participants answer some basic demographic questions. The specifics of the prediction task vary from condition to condition, and are described in detail in Section 2.1. Each participant is only allowed to participate once in the experiment.

Our 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). Participants complete the experiment using their own electronic devices (e.g. computers and tablets). For MTurk, we use standard MTurk HITs titled "Making Statistical Forecasts." The MTurk platform is commonly used in experimental studies (Kuziemko et al., 2015; D'Acunto, 2015; Cavallo et al., 2017; DellaVigna and Pope, 2017a,b). It offers a large subject pool and a more diverse sample compared to lab experiments. Prior research also finds the response quality on MTurk to be similar to other samples and to lab experiments (Casler et al., 2013; Lian et al., 2018). We also run our experiment with MIT EECS students to cross check with our MTurk results. MIT EECS students are recruited via emails sent to all EECS undergraduates, which provides a link to the experimental interface. We discuss payments and participant characteristics in more detail in Sections 2.2 and 2.3.

2.1 Experimental conditions

We conducted four rounds of experiments sequentially, which we describe below: Experiment 1 (Baseline, MTurk), Experiment 2 (Common path, MTurk), Experiment 3 (Robustness checks, MTurk), and Experiment 4 (Describe DGP, MIT EECS). Table 2 provides a summary of the experiments.

Experiment 1 (Baseline, MTurk). Experiment 1 is our baseline test. It was conducted in February 2017. The different treatments in Experiment 1 are summarized in Table 2, Panel A.

In the experiment, each participant is presented with realizations from an AR(1) process:

$$x_{t+1} = \mu + \rho x_t + \epsilon_t \tag{1}$$

Participants start with 40 past observations of the process. In each round, participants observe the new realization x_t, and are asked to predict the value of the next two realizations x_{t+1} and x_{t+2}. Figure 1 provides a screen shot of the prediction page. Specifically, a series of green dots show past realizations of the process. Participants can drag the mouse to indicate their prediction for the next realization, $F_t x_{t+1}$, in the purple bar, and indicate the following realization, $F_t x_{t+2}$, in the red bar. Participants' predictions are shown as yellow dots. We also display the prediction of x_{t+1} from the previous round $F_{t-1}x_{t+1}$ using a grey dot (participants can see it but cannot change it). After making their decisions, participants click "Make Predictions" and move on to the next round.

We focus on AR(1) processes because they are simple and not very restrictive. They are easy to learn as discussed more below in Section 3.2 (in-sample least square learning approaches full information rational expectations very closely). In addition, as Fuster et al. (2010) point out, in finite sample, ARMA processes with longer lags are difficult to tell apart from AR(1) processes statistically, and standard procedures like BIC often detect such processes as AR(1). Accordingly, they can be "observationally equivalent" to AR(1) processes to the econometrician and the forecaster. Relying on simple and clear AR(1) processes thus makes the definition of rational expectations relatively clear in this context, in particular when participants are told about the data-generating process as in our conditions with MIT undergrads (more on this later).

In this experiment, we use 6 different values of ρ: $\{0, .2, .4, .6, .8, 1\}$. The volatility of ϵ is 20. The constant here is zero: $\mu = 0$. Participants are randomly assigned to one value of ρ. Each participant is presented with a different realization of the process. There are 270 participants in total and about 30 participants per value of ρ (the randomization is not perfectly even across conditions in a finite sample).

Experiment 2 (Common path, MTurk). In Experiment 2, we study potential heterogeneity in participants' responses to the same statistical process. Thus we perform an experiment that is similar to Experiment 1, except that we use the same value of $\rho = .6$ for all participants and 10 (randomly generated) paths of realizations of the AR(1) process. Each participant is randomly assigned to one of the 10 paths. Other aspects of the experimental procedures are the same as Experiment 1. Experiment 2 was conducted in March 2017. There are 330 participants in total, with about 30 participants per path (again the randomization is not perfectly even). Treatments in Experiment 2 are summarized in Table 2, Panel B.

Experiment 3 (Robustness checks, MTurk). In Experiment 3, we modify Experiment 1 in several ways to perform robustness checks. Every participant is randomly assigned to one of these conditions. There are 875 participants in total, with roughly 35 participants per condition. Experiment 3 was conducted in June 2017. Table 2, Panel C, provides a summary of the treatments in Experiment 3.

The treatments in Experiment 3 are designed to help us implement three main tests:

1. *Well-known economic variables vs abstract process*

   In Conditions C1 to C8, we test whether participants' forecasting behavior is different when making forecasts about abstract random processes and economic variables. Specifically, we estimate the properties of four major economic variables (assuming an AR(1) process): U.S. quarterly GDP growth, monthly CPI, monthly S&P 500 stock returns, and monthly house price growth. We then use the estimated parameters to generate the random processes in the experiment. In Conditions C1 to C4, in the experimental instruction we explain that "the process you will see has the same property as [...]." In Conditions C5 to C8, we use the same random processes but only describe them as random processes (as in Experiment 1). Everything else is the same as Experiment 1. Through this design, we can examine whether participants' behavior is influenced by the "context" by comparing Conditions C1 to C4 with their counterparts in Conditions C5 to C8.

2. *Varying other AR(1) parameters than ρ: Constant and volatility*

   In addition, this experiment allows us to evaluate the effect of changing other process parameters such as the constant (μ) and the innovation volatility (σ). We use condition C9, for which $\mu = 0$, $\sigma = 20$ and $\rho = .4$ as the benchmark (the parameters are the same as those in Experiment 1 condition A3, but the advantage of C9 is that it is run at the same time as the rest of Experiment 3). Then, conditions C5-C8 discussed above (macro processes without describing economic context) can be studied in comparison to C9 as they correspond to different values of μ and σ.

3. *The term structure of expectations*

   In the remaining conditions of Experiment 3, we test the impact of asking participants to report the term structure of expectations. In conditions 10 to 13, we ask for the $t+1$ forecast only. In conditions 14 to 17, we ask for the $t+2$ forecast only. In conditions 18 to 21, we ask for the $t+1$ and $t+5$ forecasts. Finally, in conditions 22 to 25, we ask for $t+1$ and $t+2$ forecasts, but remove the grey dot that shows the $t+2$ forecast from the previous round, i.e. $F_{t-1}x_{t+1}$.

Experiment 4 (Describe DGP, MIT EECS). In Experiment 4, we study whether providing more information about the data generating process affects forecasts. To make sure participants have a good understanding of AR(1) processes, we perform this test among MIT undergraduate students in Electrical Engineering and Computer Science (EECS). Experiment 4 was conducted in March 2018. We send a recruitment email to all enrolled EECS undergraduates with a link to the experimental interface. The interface closes after 200 enrollments (which was completed in less than 12 hours). Table 2, Panel D, provides a summary of the treatments in Experiment 4.

We use AR(1) processes as in Experiment 1, with persistence $\rho = .2$ and $\rho = .6$. For each persistence, the control group is the same as Experiment 1, and the process is described as "a random process." For the treatment group, we describe the process as "a fixed and stationary AR(1) process: $x_t = \mu + \rho x_{t-1} + e_t$, with a given μ, a given ρ in the range [0,1], and e_t is an i.i.d. random shock." Thus there are 2 × 2 = 4 conditions in total. Participants are randomly assigned to one of the conditions. In the demographic section after the experiment, we make sure students understand AR(1) processes by asking them to calculate conditional expectations and variances of a standard AR(1) process.

2.2 Payments

The payments consist of fixed participation payments and incentive payments that depend on performance in the prediction task. For the incentive payments, the participant receives a score for each prediction that is a decreasing function of the forecasting error (Dwyer et al., 1993; Hey, 1994):

$$S = 100 \times \max(0, 1 - |\Delta|/\sigma) \tag{2}$$

where Δ is the difference between the prediction and the actual realization, and σ is the volatility of the noise term ϵ (20 in most conditions). For each round, and for each forecasting horizon, the score is between 0 and 100. The score is increasing with the accuracy of the forecast; if the forecast is off by too much, the score is zero. 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, and the score associated with each of the past prediction (if the actual is realized) is displayed at the bottom of the screen (see Figure 1).

The loss function defined in Equation (2) ensures that a rational participant will choose the rational expectation as an optimal forecast if there is no cost to do so (Dwyer et al., 1993; Hey, 1994). $E(1 - |x_{t+1} - F_t|/\sigma)$ is maximal for a forecast F_t equal to the 50^{th} percentile of the distribution of x_{t+1} conditional on x_t. Given that our process is symmetrical around the rational forecast, the median is equal to the mean, and the optimal forecast is therefore equal to the conditional expectation. A fully rational agent would expect to earn a total score of about 2,800 in regular conditions with 2 forecasts per round.[4]

[4] As it turns out, whether the fully rational agent knows the true ρ of the process (Full Information Rational Expectations) or predicts realizations using linear regressions (Least Square learning) does not change the expected

For experiments on MTurk (Experiments 1, 2, 3), the base payment is $1.8; the conversion ratio from the score to dollars is 600, which translates to about payments of about $5 for fully rational agents for regular conditions with 2 forecasts per round (and $2.5 for conditions in Experiment 3 with 1 forecast per round). For experiments with MIT students (Experiment 4), the base payment is $5; the conversion ratio from the score to dollars is 240, which translates to about payments of about $12 for fully rational agents. The payments are delivered in the form of Amazon gift cards via email. Table 3 shows the summary statistics of the incentive payments for all experiments separately (and within Experiment 3, splits between conditions with 1 and 2 forecasts per round). Table 3 also shows the duration of participation. The mean duration is about 15 minutes. The mean duration for each round of forecast is about 10 seconds.

Finally, the participation constraint of subjects is likely to be satisfied. For the MTurk tests, the average realized total payment (participation payment plus incentive payment) is about $5 (for a roughly 15 minute task), which is high compared to the average pay rate on MTurk. For the MIT tests, the average realized total payment is around $15. The payments are sufficiently attractive to recruit 200 EECS undergrads out of 1,291 students within 6 hours. As far as the incentive compatibility constraint is concerned, the question is more difficult as it depends on the cost of making more rational expectations. However, recent work by DellaVigna and Pope (2017b) show that participants provide high effort even when the size of the incentive payment is modest, and the power of incentives does not appear to be a primary issue in this setting.

2.3 Descriptive Statistics

Table 4 shows the demographics of participants in our experiments. In Panel A, we present basic participant characteristics. For MTurk participants, about 55% to 60% of the participants are male. Roughly 75% report they have college or graduate degrees, and the level of education is higher than that in the general US population (60% with college degrees or above) (Ryan and

score by much with 40 rounds of realizations to start with. In simulations, over 1,000 realizations of the process, we find that expected scores of the two types of rational agents differ by less than .3%. The standard deviation of the FIRE agent is, as expected however, lower by about 10%. Knowing the true ρ does not affect expected score but reduces earnings volatility.

Bauman, 2015). 40% report they have taken a statistical class. The median age is about 33, slightly lower than the general population (37) (Howden and Meyer, 2011); less than 2% of the participants are above 65. As expected, MIT EECS undergrads are much younger. They report similar levels of stock market experience, and 43% of the participants are male.

In Panel B, we present performance on standard statistical questions that we include at the end of the experiment, which include calculating the median of 8 numbers, judging whether unbalanced gender of newborns is more likely to occur at larger or smaller hospitals, and judging whether tossing fair coins is more likely to generate mixed or consecutive heads and tails (Tversky and Kahneman, 1974). Unsurprisingly, MIT undergrads did significantly better on each of these questions. About 90% of them gave correct answers to the median question and the coin toss question, versus 50% and 60% respectively for the MTurk participants. About 60% (33% among MTurk participants) gave correct answers to the more challenging question that unbalanced gender of newborns is more likely for smaller hospitals.

3 Empirical Model

This Section describes our empirical model of expectations, which nests rational, extrapolative and sticky expectations.

3.1 Main Components

We define here notations for expectations. Consider a random variable x_t that an agent is trying to forecast. We denote by $F_{t-k}x_{t+1}$ the subjective forecast of x_{t+1}, k periods ahead of time t. This is the object that we measure through our experiments. Then, denote by $E_{t-k}x_{t+1}$ the rational expectation of x_{t+1} as of time $t-k$. We discuss in Section 3.2 how this rational expectation is measured.

We specify potential biases in the subjective forecast $F_{t-k}x_{t+1}$ using a term for extrapolation and a term for stickiness. Below we first describe how the two parts are defined. We then combine them in a single specification.

We model extrapolative expectations F^e as:

$$F^e_{t-k}x_{t+1} = E_{t-k}x_{t+1} + \gamma(x_{t-k} - E_{t-k-1}x_{t-k}) \tag{3}$$

where γ captures the strength of extrapolation. This specification nests rational expectations as a special case ($\gamma = 0$). Extrapolative individuals ($\gamma > 0$) react too much to recent innovations, and perceive that the shock will persist more than it actually does. Note that if x_t has a deterministic trend and no deviations from the trend (as in: $x_t = x_{t-1} + g$), subjective expectations would be rational with this specification. In other words, only unexpected positive deviation from the trend will generate over-optimistic expectations and vice versa. This specification is similar to Bordalo et al. (2018c) and Bordalo et al. (2018a); a minor difference in functional form is that we use $x_{t-k} - E_{t-k-1}x_{t-k}$ to capture the shock while Bordalo et al. (2018c) use $E_{t-k}x_{t+1-k} - E_{t-k-1}x_{t+1-k}$ to capture the shock. Thus, in their model, even deterministic trends with i.i.d. shocks (as in: $x_t = x_{t-1} + g + u_t$) do not generate biased forecasts.

We model sticky expectations F^s using the recursive formulation:

$$F^s_{t-k}x_{t+1} = (1 - \lambda)E_{t-k}x_{t+1} + \lambda F^s_{t-k1}x_{t+1} \tag{4}$$

where $\lambda \in [0, 1]$ measures the degree of stickiness. $\lambda = 0$ corresponds to fully rational expectations. When $\lambda > 0$, the agent's current beliefs carry over her past beliefs. Sticky expectations capture the phenomenon that agents' current forecasts have a tendency to anchor on previous beliefs. This formulation resembles the specification in Coibion and Gorodnichenko (2015). A key difference is that Coibion and Gorodnichenko (2015) use the specification to describe the evolution of consensus (mean) forecasts, which may arise from infrequent updating by some forecasters (Mankiw and Reis, 2002) or from heterogeneous noisy private information (Woodford, 2003). In our specification, sticky expectations are allowed to occur at the *individual* level, as in Bouchaud et al. (2018). Estimating this model requires that we measure forecasts at two different horizons, which are available in our experiment.

The empirical specification that we use in this paper combines the two formulations, and describes the subjective forecasts as follows:

$$F_{t-k}x_{t+1} = (1-\lambda)E_{t-k}x_{t+1} + \lambda F_{t-k-1}x_{t+1} + \gamma(x_{t-k} - E_{t-k-1}x_{t-k}) \tag{5}$$

After rearranging, we can express expected forecast errors for $k = 0$ as:

$$F_t x_{t+1} - E_t x_{t+1} = \underbrace{\lambda\left(F_{t-1}x_{t+1} - E_t x_{t+1}\right)}_{\text{stickiness}} + \underbrace{\gamma(x_t - E_{t-1}x_t)}_{\text{extrapolation}} \tag{6}$$

In this specification, the individual forecaster can be both extrapolative $\gamma > 0$ and sticky $\lambda > 0$. These two effects can be estimated by regressing the expected expectation error $F_t x_{t+1} - E_t x_{t+1}$ on the distance of past period expectation to current period rational expectation, $F_{t-1}x_{t+1} - E_t x_{t+1}$, which captures stickiness, and past period innovation $x_t - E_{t-1}x_t$, which captures extrapolation. Intuitively, if expectation errors can be forecasted using previous period errors, this is a sign of stickiness (errors persist), and $\lambda > 0$. If expectation errors can be forecasted using past innovation, this is a sign of extrapolation, and $\gamma > 0$. If forecasts are rational, then $\lambda = \gamma = 0$.

The combination of stickiness and extrapolation plays an interesting role. In particular, stickiness can propagate past biases. Without the stickiness component, extrapolation adds a distortion each period onto the rational expectations in that period (i.e. subjective belief exaggerates the impact of the recent innovation), as in Bordalo et al. (2018c), and past biases do not matter. With the stickiness component, extrapolation adds a distortion each period onto a starting point that anchors partly on previous beliefs. As a result, past biases can persist and continue to affect future forecasts (we return to this issue of bias persistence in Section 4.2).

3.2 Measuring Rational Expectations

To estimate our econometric specification, we need to compute the rational expectation of the agent, $E_{t-k}x_{t+1}$. We use two different measures, which we describe here.

In all of our regressions (unless otherwise specified), we use Least Square learning to define our baseline measure of rational expectations. Given that the participant does not know the data-generating process, it is reasonable to define rational expectations in this set-up through rolling regressions. We follow Evans and Honkapohja (2001) and assume rational participants would use "least square learning" to estimate the following linear model:

$$\widehat{E}_{t-k}x_t = a_{t-k} + \sum_{h=0}^{h=n} b_{i,t-k}x_{t-k-h}$$

In period $t-k$, the participant forecasts x_t using lagged values from x_{t-k-n} up to x_{t-k}. Parameters a_{t-k} and $b_{h,t-k}$ are estimated using OLS and past realizations of x_t until x_{t-k}. In the paper, we set $n = 3$ to be conservative, but our results are not sensitive to this threshold. This definition of LS learning-based expectations is our *baseline* measure of rational expectations.

In the main results in Section 4, we also present specifications with full information rational expectation (FIRE):

$$E^{FI}_{t-k}x_t = \rho^k x_t$$

FIRE is a strong assumption as it assumes that agents know the true data-generating process. Nevertheless, since it is commonly used in economic models, we report results using FIRE as well.

As it turns out, both definitions of rational expectations give nearly identical results. We report most results using the LS formulation of rational expectations because it seems to be the more natural choice, but using FIRE gives the same output. The reason is that LS learning expectations are already quite close to FIRE in our setting (stationary AR(1) processes). Regressing FIRE on LS expectations, we find a slope coefficient of .86 and a R^2 of .84. The two definitions of rational expectations are close to one another because the process is simple (an AR(1)), and participants are provided with enough data points to learn from (40 data points to begin with). Panel A of Figure 2 shows that the mean squared difference between these two expectations does not decrease very fast during the time the experiment takes place (we discuss Panel B below). This is mostly because the experiment starts with 40 historical observations, so the estimated model is already quite precise when participants make their first forecast. As is well known, such similarity may not hold for a more complex data generating process, e.g. ARMA processes with long lags (Fuster et al., 2010) or processes with non-linear terms. Note that the LS learning expectation definition implicitly assumes that participants start with a prior that the data generating process is linear. We return to this issue later when discussing Experiment 4 (with MIT undergrads who are told that the process is a fixed, stationary AR(1)).

Finally, in Figure 3 we show the distribution of % score loss of subjective forecasts relative to LS rationality in Experiment 1. For each participant, we compute S^*, the score this participant would have obtained using LS rational expectations. We then compute $S/S^* - 1$, the score loss relative to LS rational expectations, and Figure 3 shows the distribution. 217 out of 271 participants have a negative loss, and the average loss is about 26%. Some participants end up lucky because, in spite of wrong forecasts, ex post realizations ended up going in the right direction.

4 Empirical Results

4.1 Main Results

We now turn to our main results. As discussed in Section 3, we run the following regression using the forecast data we collect. The baseline results use data from Experiment 1 (Baseline, MTurk), pooling all participants for which the persistence parameter $\rho \in \{0, .2, .4, .6, .8, 1\}$. For individual i at date t:

$$F_t^i x_{it+1} - E_t x_{it+1} = \lambda \left(F_{t-1}^i x_{it+1} - E_t x_{it+1} \right) + \gamma (x_{it} - E_{t-1} x_{it}) + u_{it+1} \tag{7}$$

where the rational expectation $E_t x_{t+1}$ is in general measured using least square learning $\hat{E}_{t-k} x_t$, as discussed in Section 3.2, except when noted (columns labeled “FIRE"). We use OLS and cluster standard errors u_{it+1} at the individual level.

Results are reported in Table 5. Column (1) assumes no extrapolation ($\gamma = 0$). Expectations appear to be sticky with a coefficient $\lambda = .16$, strongly significant statistically (t-stat of 3.8). An econometrician ignoring potential extrapolation would thus infer that expectations are "16%" sticky and "84%" rational. Note that this coefficient is in the ballpark of estimates of expectation stickiness in the literature (Coibion and Gorodnichenko, 2015; Bouchaud et al., 2018). Column (2) performs the opposite exercise, assuming pure extrapolation, and indeed finds evidence of extrapolation, with $\gamma = .36$, significant with a t-stat of 18. The two components are included together in column (3), which is our main specification. Compared to columns (1) and (2), both γ and λ increase and are both very significant. Column (5) confirms our main finding using the Full Information Rational Expectation (FIRE) instead of LS learning RE (thus using perfectly informed participants as the rational benchmark). Estimates are very similar, which is consistent with our earlier discussion that the two measures of rational expectations are highly correlated. Overall, across columns (3)-(5), λ hovers between .25 and .26; γ hovers between .44 and .46. We provide more discussion about magnitude in Section 4.2.

Finally, in columns (6) and (7), we investigate the possibility that learning takes place and reduces systematic errors over time. We do not find much evidence of learning during the 40 periods of our test. We split the sample between the first 20 and the last 20 rounds of testing. If learning takes place, we should see a reduction in the estimated λ and γ in the last 20 rounds, which we do not observe in the data. Both stickiness λ and over-reaction γ increase a bit, but none of these changes are statistically significant.

Another way to explore learning is by computing the mean squared difference between subjective forecasts and rational forecasts. We show this statistic in Figure 2, Panel B, where we show the square root of the mean squared difference between the observed forecast $F_{t-1}x_t$ and the LS forecast $\widehat{E}_{t-1}x_t$. We plot this number as a function of the round of observation in Panel B. It does not show any evidence of decreasing over time, consistently with the idea that little convergence towards rational expectations is taking place.

In Internet Appendix A.1, we also study the possibility that participants have regime-switching priors (Barberis et al., 1998; Rabin, 2002; Rabin and Vayanos, 2010), and react to signals in a path-dependent manner. For instance, after observing long streaks in the past, the forecaster may be more likely to predict continuation. We do not find strong evidence of such behavior, as shown in Table A.1 of the Appendix. One possibility is that previous findings of regime-switching beliefs generally come from settings with binary outcomes (Asparouhova et al., 2009) or pre-defined paths of continuous outcomes (Bloomfield and Hales, 2002), while we allow for unlimited possible realizations of continuous outcomes. With continuous outcomes, whether a given pattern represents streak can be less obvious. In addition, as discussed later in Section 5.6, we obtain very similar results when we explain to participants that the process is a fixed, stationary AR(1) in the test with MIT EECS students. In that case, we explicitly rule out regime-shifting priors, and find that our results are unchanged.

4.2 Quantifying the Results

We further analyze the quantitative impact of stickiness and extrapolation. We start from the equivalent formulation:

$$F_t x_{t+1} = (1-\lambda)\sum_{k=0}^{\infty} \lambda^k E_{t-k} x_{t+1} + \gamma \sum_{k=0}^{\infty} \lambda^k \left(x_{t-k} - E_{t-k-1} x_{t-k}\right) \tag{8}$$

This non-recursive formulation is equivalent to our main recursive model in Equation (7). We estimate it directly in Appendix A.2 and estimates $\lambda \approx .22$ and $\gamma \approx .46$ are similar to the ones we obtain in Table 5. These estimates are stable when introducing 2, 3, or 4 lags, and when splitting data into the first and last 20 rounds of testing.

As Equation (8) points out, stickiness has two opposite effects on expectations formation. The first effect shows up in the first term of the RHS of Equation (8): It is the standard under-reaction intuition emphasized in the existing literature (Coibion and Gorodnichenko, 2015; Bouchaud et al., 2018). A bigger λ means that forecasts are more “stuck” to previous expectations. The second effect shows up in the second term: The forecaster’s propensity to extrapolate is also itself sticky.

All past over-reactions have an effect on current forecasts, as in Barberis et al. (2015). Another way to say this is that λ governs the window length for measuring extrapolation: A higher λ means that the agent looks further back in time to estimate the trend that determines extrapolative beliefs.

One simple way to see and quantify these two opposite effects at work is to rearrange and approximate the lag formulation (8) into the following expression:

$$\begin{aligned} F_t x_{t+1} - E_t x_{t+1} &= \sum_{k=0}^{\infty} \underbrace{\lambda^k (\gamma - \lambda \rho^{k+1})}_{a_k} \epsilon_{t-k} \\ &\approx (\gamma - \lambda\rho)\epsilon_t + \lambda\gamma\epsilon_{t-1} \end{aligned}$$

where we neglect for pedagogical purposes all terms in λ^2 and more (given the higher order terms are very small). The first term on the RHS contains the net reaction to the latest innovation. It is positive when $\gamma > \lambda\rho$, which means that there is net over-reaction to news. This condition is satisfied in all of our specifications. The second term contains the pure effect of past over-reaction. In this term, stickiness does not produce under-reaction, but induces persistence in past over-reaction. It is large when $\lambda\gamma$ is large, i.e. over-reaction is significant but stickiness is big too.

For the average forecaster, it is clear from the above equation that (1) over-reaction dominates quantitatively and (2) it persists but not much beyond two rounds. To illustrate this, we show in Figure 4 the impulse response for the case where $\rho = .6$. We represent rational expectations along with simulated forecast using formula (8) for $\lambda = .26$, $\gamma = .46$. The impulse response shows that our forecasters over-react to the impulse when compared to rational forecasts. The first period excess response is $\gamma - \lambda\rho = .3$. The second period response is also non-negligible: $\lambda\gamma = .11$: over-reaction persists by about a third after one period.

This persistence in over-reaction is not only important for the average forecast, but also important for explaining the cross-section of forecasts. Table 5 shows that adding the stickiness component doubles the cross-sectional R^2 from 10% to 20%.

4.3 Heterogeneity

Previous results show that our model in Equation (7) can capture the average behavior of expectations in the data. Below we investigate the heterogeneity among individuals. Specifically, in our main model we now allow the coefficients λ and γ to vary across individuals:

$$F_t^i x_{it+1} - E_t x_{it+1} = \lambda_i \left(F_{t-1}^i x_{it+1} - E_t x_{it+1} \right) + \gamma_i (x_{it} - E_{t-1} x_{it}) + u_{it+1} \tag{9}$$

We run one regression for each participant. We still use the baseline sample from Experiment 1.

In Figure 5, we show the distributions of the stickiness and extrapolation parameters. Two messages emerge. First, the null hypothesis of (LS based) rational expectations ($\lambda = \gamma = 0$) is rejected at 5% for 237 out of 270 subjects. Second, there is significant dispersion, but a lot of it is due to the small number of observations (39 rounds per each individual i) used to estimate each parameter. To assess the heterogeneity separately from estimation noise, for each individual, we compute the p-value of a test of the null that $\lambda_i = \lambda$, taking into account the fact that both numbers are estimated (we run the two regressions using the SURE approach). The stickiness parameter differs from average at 5% for 80 out of 270 participants. The extrapolation parameter differs from average for 83 participants. Overall, we cannot reject that about half (131) of the individuals behave like in the average model. Expectations formation in our data appears reasonably well described, even in the cross-section, by our empirical model.

We also study the cross-sectional properties of the individual parameters. To do so, we regress individual λ_i and γ_i on various participant characteristics. Consistently with our robustness checks in Tables 7 and 8 below, we do not find any significant and consistent relation between either of the two parameters and socio-demographics, measures of statistical literacy, or experimental settings. λ_i and γ_i do not vary systematically across these groups. The only cross-sectional relation that emerges is the negative correlation between λ_i and γ_i which is equal to $-.2$ and significant at 1%. Hence, sticky subjects tend to extrapolate less.

4.4 Individual-level vs Consensus Forecast

Given that we do not find substantial heterogeneity across individuals, we further analyze if our empirical model of expectations formation does well to explain aggregate expectations in addition to individual expectations. The model estimated in Table 5 has an R^2 of .20, which suggests that the error term u_{it+1} in Equation (7) is reasonably volatile. This means that individual expectations do contain noise. The consensus forecast, however, may be easier to predict if the individual noise is idiosyncratic.

To analyze consensus forecasts, we need to ensure that participants see the same realizations of the AR(1) process. We do this in Experiment 2 (Common path, MTurk). As discussed in Section 2.1, in this experiment we use AR(1) processes with $\rho = .6$. We then randomly assign 330 subjects into 10 different paths of this process. The 10 paths are decided by randomly chosen seeds. For each path, there is about 30 participants seeing the same realizations. We calculate the average expectations for each path, and test how well our baseline model does for the consensus forecasts. We estimate the following regression:

$$\bar{F}_t x_{ct+1} - E_t x_{ct+1} = \lambda \left(\bar{F}_{t-1} x_{ct+1} - E_t x_{ct+1} \right) + \gamma (x_{ct} - E_{t-1} x_{ct}) + v_{ct}$$

for condition (path) c at round t. $\bar{F}_t x_{ct+1}$ is the *average* prediction across subjects in condition c at round t for realization x_{t+1}. Given we are using average expectations, the panel on which we run these regressions is smaller than that in Table 5. We only observe 40 rounds across 10 different conditions, and therefore have at most 400 observations in total.

We report the results of this regression in Table 6 using the same structure as in Table 5: with the stickiness and extrapolation components separately first then together, with LS RE and full information RE, for the first and last 20 periods separately. Several findings emerge. First, the coefficients obtained in this setting are very similar to the coefficients obtained in Table 5. Comparing columns (3) of both Tables, we find $\lambda = .29$ (vs .25) and $\gamma = .47$ (vs .46). Second, the R^2 of this regression (.58) is much higher than in our cross-sectional specification (.19). Thus, a big part of the error term u_{it} in the individual expectation model appears to be idiosyncratic noise that cancel out in aggregate. Overall, our model does a very good job at explaining the average expectation formation process.[5] Third, using LS rational expectations vs. FIRE makes no real difference at the aggregate level: The model with FIRE has the same R^2 and similar overreaction and stickiness parameters. Finally, the model works better for the last 20 rounds than in the first 20 rounds. This last finding is consistent with the idea that in-sample LS learning is more realistic.

5 Robustness

In this Section, we investigate the robustness of our results to participant characteristics and changes in the experimental setting.

5.1 Demographics

Table 7 offers further evidence that our estimates of λ and γ are very stable across subpopulations. We use the baseline sample from Experiment 1; results in other samples are very similar. In Panel A, we split the sample by demographics: Gender (columns (1)-(2)), Age (columns (3)-(4)) and Education (columns (5)-(6)). In Panel B, we split the sample of participants by response to basic questions designed to test the statistical skill of participants as discussed in Section 2.3. In columns (1)-(2), we focus on the "coin toss" question, designed to test if participants understand the notion of statistical independence. In columns (3)-(4), we look at answers to a question designed to see if participants can correctly calculate the median. In columns (5)-(6), we split participants into those who answered right or wrong to the "hospital" questions, which tests if people understand the law of large numbers.

In all the subsamples, the stickiness estimate is strongly statistically significant and hovers between .21 and .32. The extrapolation parameter is even more homogeneous and hovers between

[5] Coibion and Gorodnichenko (2015) and Bordalo et al. (2018b) show that *when informational frictions are present* (e.g. agents receive heterogeneous private information), then the behavior of consensus forecasts can be different from the behavior of individual-level forecasts. We design a simple setting with little informational frictions, and we find that such differences do not appear in this case.

.44 and .47. Interestingly, measures of statistical skill based on standard test questions have little effect on the estimates. Overall, none of the 6 sample splits tried here generates a significant difference in λ or γ. Behavior appears surprisingly stable across different demographic groups.

5.2 Varying Process Persistence

We then test the stability of the results for AR(1) processes with different levels of persistence ρ, using data from Experiment 1 where participants are randomly assigned to conditions with different values of ρ. Table 8, Panel A, reports the estimation of Equation (7) for each value of ρ between 0 and 1. We first focus on stationary processes, i.e. processes for which ρ is between 0 and .8. For these processes, the model turns out to be remarkably stable. The stickiness coefficient lies between .12 and .24. The same result holds for the extrapolation parameter, which lies between .41 and .48. The R^2 is also very stable, which hovers between .15 and .20 across different values of ρ. The qualitative findings remain the same for the condition where $\rho = 1$, but both stickiness and over-reaction increase sizably. However, this condition does not generate significantly different parameters. The null that γ is stable is still not rejected (p-value of .53). The null that stickiness λ is stable is also not rejected, though with a lower p-value of .15. It thus looks like non-stationary processes are harder to cope with.

Other specifications of expectations formation perform less well across different values of persistence ρ. Table 8, Panel B shows that if we leave out the stickiness component, the results for extrapolation are less stable. In Panel C, we estimate the pure "sticky model," which appears to be imprecise for both $\rho = .4$ and $\rho = .8$. Panel D turns to a "backward-looking" version of our model. This backward-looking version replaces the RE terms with x_t, so that the specification becomes:

$$F_t^i x_{it+1} - x_{it} = \lambda \left(F_{t-1}^i x_{it+1} - x_{it} \right) + \gamma (x_{it} - x_{it-1}) + u_{it+1} \tag{10}$$

In this case, the forecaster relies purely on historical realizations, and does not make adjustments based on the properties of the true process (no attempt to account for the imperfect persistence).

This model appears to fit the data less well than our baseline specification shown in Panel A. First, the R^2 is lower by one third or more in all conditions corresponding to stationary processes. Second, the extrapolation parameter is very unstable, going from -.13 to .58 across all stationary processes. The test of equality is strongly rejected. Finally, the results for $\rho = 1$ are similar to our baseline, but this is hardwired since with a random walk $E_t x_{it+1} = x_{it}$. With such a process, “backward-looking" expectations are mechanically rational: It is impossible to distinguish rational expectations from myopic beliefs. Varying ρ from 0 to 1 is indeed one of the contributions of our design, as most existing studies work with random walks (see Table 1), where it may be difficult to distinguish between rational vs. irrational updating.

5.3 Varying Other Parameters

In this Section, we vary other process parameters than just the persistence. We use results from Experiment 3 (Robustness checks, MTurk). In the baseline test in Table 8, we use AR(1) processes with $\mu = Ex_t = 0$ and $\sigma = 20$. This is still the case in condition C9 in Experiment 3, with $\rho = .4$ as shown in Table 2. In conditions C5-C8, we change μ and σ as well, as reported in Table 2; everything else is the same across these conditions (and is the same as in the baseline test of Experiment 1). In these conditions, volatility of the shock varies widely, from .23 to 20. The constant μ goes from 0 to .55.

Table 9 reports the results. Column (1) shows condition C9, for which $\mu = 0$, $\sigma = 20$ and $\rho = .4$. Results are generally similar to those in Table 5. Columns (2)-(4) show results in conditions C5-C8. While the parameters of the underlying AR(1) processes differ substantially, estimates of our model parameters (λ, γ) are quite similar. The p-value of tests of equality between each of the four conditions and the baseline in column (1) are presented in the bottom panel of the Table. The null hypothesis of equality is never rejected.

5.4 Reporting the Term Structure of Expectations

We also investigate the effect of reporting different term structures of expectations. A key dimension of our experimental setting is that we ask participants to report long-term expectations. This may strengthen anchoring—because they are asked to report long-term expectations ($F_t x_{t+2}$), subjects may have a propensity to stick to their long-term expectations. We investigate this question using several conditions, and present the results in Table 10. Our tests suggest that stickiness is indeed affected by the reporting of long-term expectations. However, the extrapolation coefficient is very robust across all conditions.

First, in Table 10 columns (1)-(2), we ask if the presentation of past long-term expectations affects forecasting behavior. In our baseline experimental setting, in round t, we use a gray dot to represent previous long-term expectations $F_{t-1} x_{t+1}$. In other words, when participants make forecasts $F_t x_{t+1}$ and $F_t x_{t+2}$, they can see via the gray dot the prediction of x_{t+1} made in the previous round (this information disappears after round t). The gray dot helps remember past forecasts, which may reinforce anchoring of expectations. To study the influence of the grey dot, we randomly assign we participants in Experiment 3 into two types of treatments that both asking for short- and long-term forecasts (as in the baseline test), but in one of the treatments we remove the gray dot.

We run our main specification (7) and report the results in Table 10: column (1) corresponds to a condition with gray dots where $\mu = 0$, $\rho = .4$, and $\sigma = 20$, as in our baseline analysis. Column (2) corresponds to a condition where the parameters are the same, but there is no gray dot. We test the equality of coefficients in the bottom panel. The extrapolation coefficient γ is not statistically different across the two conditions (.45 without gray dot against .48 in the baseline). The stickiness coefficient λ is however significantly higher with the gray dot (p-value of .03). It is equal to .24*** in the baseline condition versus .08 in the condition without gray dot. Thus, the presence of the gray dot tends to make expectations stickier. Meanwhile, extrapolation is unchanged by the absence of the gray dot.

Second, we ask if the mere fact of reporting long-term expectations tends to make short-term expectation stickier. We implement these tests in columns (3)-(5). In column (3), we show the baseline condition where $\mu = 0$, $\rho = .4$, and $\sigma = 20$. In column (4), we report the results of a condition where the process parameters are the same but subjects are only required to provide short-term expectations $F_t x_{t+1}$ and not long-term ones $F_t x_{t+2}$ (C11 in Table 2). In column (5), on the contrary, we analyze a condition where subjects report short-term expectations and very long-term ones $F_t x_{t+5}$ (C19 in Table 2).

To compare these three conditions, we cannot run our main specification (7) since it requires both $F_t x_{t+1}$ and lagged $F_t x_{t+2}$, which we do not have in these two alternative conditions. Instead, we run the lagged equivalent of (7):

$$F_t^i x_{t+1} = (1-\lambda) \sum_{k=0}^{2} \lambda^k E_{t-k}^i x_{it+1} + \gamma \sum_{k=0}^{2} \lambda^k \left(x_{it-k} - E_{t-k-1}^i x_{it-k} \right) + \eta_{it} \tag{11}$$

which is the same equation as in (8) limited to three lags, since coefficients are generally negligible after 2 lags. Note that the coefficient on $E_t^i x_{it+1}$ should be equal to $1 - \lambda$, while the coefficient on $x_{it-k} - E_{t-k-1}^i x_{it-k}$ should be interpreted as an estimate of γ. The above regression is run in columns (3)-(5) for each of the three conditions, and in the bottom panel we test equality of coefficients on $E_t^i x_{it+1}$ and $x_{it-k} - E_{t-k-1}^i x_{it-k}$ (we only focus on the first lags here in order to conserve space).

The main result here is that both coefficients are similar across the three conditions. The estimate of λ is not statistically different; the estimate of γ is a bit smaller in the second and third conditions, but has the same order of magnitude Indeed, asking participants to make long-term forecasts tends to make them slightly stickier – probably because they remember them – but the difference is not statistically significant. The evidence on extrapolation is even noisier. Overall, asking participants to report different term-structures of expectations does not seem to affect our estimates of λ and γ.

Third, we also ask if eliciting short-term expectations $F_t x_{t+1}$ affects the reporting of long-term expectation $F_t x_{t+2}$. We do this in columns (6) and (7), where we compare the baseline with a condition where participants only report $F_t x_{t+2}$ (C15 in Table 2). Again the process parameters in both conditions are set at $\mu = 0$, $\rho = .4$ and $\sigma = 20$. Like in the previous test, since we only have one expectation rather than two, we need to use the lag formulation of our model, except that now we seek to explain $F_t x_{t+2}$ (which is present in both conditions) and not $F_t x_{t+1}$. The extension of Equation (8) to this case yields:

$$F^i_{t-1} x_{t+1} = (1-\lambda) \sum_{k=0}^{1} \lambda^k E^i_{t-1-k} x_{it+1} + \gamma \sum_{k=0}^{1} \lambda^k \left(x_{it-k} - E^i_{t-1-k} x_{it-k} \right) + \eta_{it} \tag{12}$$

where the coefficient on $E^i_{t-1} x_{it+1}$ is equal to $1-\lambda$ and the coefficient on $x_{it} - E^i_{t-1} x_{it}$ is equal to γ. We run the regression separately for the two conditions in columns (6) and (7). We find that both coefficients are similar across the two settings. Long-term expectations do not seem to be too affected by short-term expectation reporting. The stickiness coefficient is marginally affected, in the direction of long-term expectations being stickier when short-term ones are reported, but the p-value is high (.37).

5.5 Economic Variables vs Abstract Processes

Experiment 3 also studies if subjects behave differently when they forecast an abstract process vs. a process associated with economic variables. We focus on four main variables: U.S. quarterly GDP growth, monthly CPI inflation, monthly S&P 500 returns and monthly house price growth. For each of these variables, we first estimate the process as an AR(1) process, which we then simulate for each participants (each participant receives a different draw of realized innovation). We then randomly assign subjects to two types of conditions. The first type of conditions are the same as the baseline test (all descriptions are the same, except the parameters of the processes are based on the economic variables); they are conditions C5-C8 in Experiment 3 shown in Table 2 (and used above in Section 5.3). The second type of conditions add to this a verbal description of the process at the beginning of the experimental instructions: "The process you will see has the same property as quarterly US real GDP growth in the last three decades" (for the GDP growth time series); they correspond to conditions C1-C4 in Experiment 3 shown in Table 2.

We estimate our main specification separately in each condition in Table 11. For each of the four economic variables (GDP, inflation, stock market and housing market returns), we compare the conditions where we mention the associated economic variable to the conditions where we just describe the process as a "random statistical process". We test equality of estimated λ and γ, and provide p-values in the bottom two lines. For all four variables, we cannot reject the null hypothesis that results across different conditions are the same. Our results echo previous findings by Frydman and Nave (2016) that participants display similar extrapolative biases in perceptual tasks and economic decisions. The fundamental biases appear stable across different contexts.

5.6 Information about Data Generating Process

Finally, we study whether providing information about the data generating process affects behavior. As mentioned in Section 2.1, to make sure participants have a good understanding of information about the data generating process, we perform this test among MIT undergraduate students in Electrical Engineering and Computer Science (EECS). We focus on this population because we can ensure that a reasonably large fraction of participants understand what an AR(1) process means (which we verify in the demographic section at the end of the experiment, as discussed more below).

We use AR(1) processes with $\rho = .2$ and $\rho = .6$, and randomly assign participants into two types of conditions as shown in Table 2. The first type of condition (control) is the same as our baseline test. The second type of condition (treatment) explicitly describes 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." This detailed description is provided at the beginning of the experiment. At the end of the experiment, we use test questions to verify that participants indeed understand this formulation. In these test questions, we ask them to calculate the conditional mean of $x_t = .5x_{t-1} + e_t$ at a one- and a two-period horizon.

Table 12 provides the results. In both the treatment and control groups, the estimates of λ and γ are very similar to our results above obtained from the MTurk population. We find no significant difference between the treatment and the control. This is true whether we take all MIT EECS participants together (columns (1) and (2)), or just those who have a good understanding of AR1 in the test questions. Also, there is no significant difference in the treatment condition between participants who answer the test questions well (.53) and all treated participants (.51). Overall, we do not find that providing information that the actual process is AR(1) has an impact. This finding is broadly in line with previous studies (Kahneman and Tversky, 1973; Tversky and Kahneman, 1974; Bloomfield and Hales, 2002; Frydman and Nave, 2016), which find that explicit information about the underlying processes (fair coin toss, i.i.d. process, random walk) does not eliminate biases. Our data also suggest that the biases are rather stable; these innate biases are significant even with information about the data generating process.

6 Conclusion

In this paper, we conduct a large-scale experiment to investigate expectations formation. At both the individual and the aggregate level, we find strong evidence of extrapolative biases and of forecast stickiness. Extrapolation is quantitatively more important, while stickiness can further propagate past extrapolative biases. Stickiness is in general stable, and stronger when participants are reminded of their past forecasts, while extrapolation does not seem to vary across all conditions. Our findings are very stable across different demographic groups and variations of experimental settings. Overall, expectations in our setting deviate significantly from the rational benchmark, and such deviations exhibit clear patterns.

References

Abarbanell, Jeffery S, and Victor L Bernard, 1992, Tests of analysts' overreaction/underreaction to earnings information as an explanation for anomalous stock price behavior, *Journal of Finance* 47, 1181–1207.

Amromin, Gene, and Steven A Sharpe, 2013, From the horse's mouth: Economic conditions and investor expectations of risk and return, *Management Science* 60, 845–866.

Asparouhova, Elena, Michael Hertzel, and Michael Lemmon, 2009, Inference from streaks in random outcomes: Experimental evidence on beliefs in regime shifting and the law of small numbers, *Management Science* 55, 1766–1782.

Assenza, Tiziana, Te Bao, Cars Hommes, and Domenico Massaro, 2014, Experiments on expectations in macroeconomics and finance, *Research in Experimental Economics* 17, 11–70.

Barberis, Nicholas, Robin Greenwood, Lawrence Jin, and Andrei Shleifer, 2015, X-capm: An extrapolative capital asset pricing model, *Journal of Financial Economics* 115, 1–24.

Barberis, Nicholas, Andrei Shleifer, and Robert Vishny, 1998, A model of investor sentiment, *Journal of Financial Economics* 49, 307–343.

Barrero, Jose Maria, 2018, The micro and macro implications of managers' beliefs, Working paper.

Beshears, John, James J Choi, Andreas Fuster, David Laibson, and Brigitte C Madrian, 2013, What goes up must come down? Experimental evidence on intuitive forecasting, *American economic review* 103, 570–574.

Bloomfield, Robert, and Jeffrey Hales, 2002, Predicting the next step of a random walk: Experimental evidence of regime-shifting beliefs, *Journal of Financial Economics* 65, 397–414.

Bordalo, Pedro, Nicola Gennaioli, Rafael La Porta, and Andrei Shleifer, 2018a, Diagnostic expectation and stock returns, Working paper.

Bordalo, Pedro, Nicola Gennaioli, Yueran Ma, and Andrei Shleifer, 2018b, Over-reaction in macroeconomic expectations, Working paper.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer, 2018c, Diagnostic expectations and credit cycles, *Journal of Finance* 73, 199–227.

Bouchaud, Jean-Philippe, Philipp Krüger, Augustin Landier, and David Thesmar, 2018, Sticky expectations and the profitability anomaly, *Forthcoming Journal of Finance* .

Casler, Krista, Lydia Bickel, and Elizabeth Hackett, 2013, Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing, *Computers in Human Behavior* 29, 2156–2160.

Cavallo, Alberto, Guillermo Cruces, and Ricardo Perez-Truglia, 2017, Inflation expectations, learning, and supermarket prices: Evidence from survey experiments, *American Economic Journal: Macroeconomics* 9, 1–35.

Coibion, Olivier, and Yuriy Gorodnichenko, 2012, What can survey forecasts tell us about information rigidities?, *Journal of Political Economy* 120, 116–159.

Coibion, Olivier, and Yuriy Gorodnichenko, 2015, Information rigidity and the expectations formation process: A simple framework and new facts, *American Economic Review* 105, 2644–78.

D'Acunto, Francesco, 2015, Identity, overconfidence, and investment decisions, Working paper.

Debondt, Werner, and Richard Thaler, 1985, Does the stock market overreact?, *Journal of Finance* 40, 793–805.

DellaVigna, Stefano, and Devin Pope, 2017a, Predicting experimental results: who knows what?, *Forthcoming Journal of Political Economy* .

DellaVigna, Stefano, and Devin Pope, 2017b, What motivates effort? Evidence and expert forecasts, *Review of Economic Studies* 85, 1029–1069.

Dwyer, Gerald, Arlington Williams, Raymond Battalio, and Timothy Mason, 1993, Tests of rational expectations in a stark setting, *Economic Journal* .

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

Frydman, Cary, and Gideon Nave, 2016, Extrapolative beliefs in perceptual and economic decisions: Evidence of a common mechanism, *Management Science* 63, 2340–2352.

Fuster, Andreas, David Laibson, and Brock Mendel, 2010, Natural expectations and macroeconomic fluctuations, *Journal of Economic Perspectives* 24, 67–84.

Gennaioli, Nicola, Yueran Ma, and Andrei Shleifer, 2016, Expectations and investment, *NBER Macroeconomics Annual* 30, 379–431.

Greenwood, Robin, and Andrei Shleifer, 2014, Expectations of returns and expected returns, *Review of Financial Studies* 27, 714–746.

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

Howden, Lindsay M., and Julie A. Meyer, 2011, Age and sex composition: 2010, US census bureau report, US Census Bureau.

Kahneman, Daniel, and Amos Tversky, 1973, On the psychology of prediction, *Psychological review* 80, 237.

Kuziemko, Ilyana, Michael I Norton, Emmanuel Saez, and Stefanie Stantcheva, 2015, How elastic are preferences for redistribution? Evidence from randomized survey experiments, *American Economic Review* 105, 1478–1508.

Lian, Chen, Yueran Ma, and Carmen Wang, 2018, Low interest rates and risk taking: Evidence from individual investment decisions, *Forthcoming Review of Financial Studies* .

Ma, Yueran, David Thesmar, and David Sraer, 2018, Do managerial forecasting biases matter?, Working paper.

Mankiw, Gregory, and Ricardo Reis, 2002, Sticky information versus sticky prices: A proposal to replace the New Keynesian Philips Curve, *Quarterly Journal of Economics* 117, 1295–1328.

Nagel, Stefan, and Zhengyang Xu, 2018, Asset pricing with fading memory, Working paper.

Rabin, Matthew, 2002, Inference by believers in the law of small numbers, *Quarterly Journal of Economics* 117, 775–816.

Rabin, Matthew, and Dimitri Vayanos, 2010, The gambler's and hot-hand fallacies: Theory and applications, *Review of Economic Studies* 77, 730–778.

Ryan, Camille L, and Kurt Bauman, 2015, Educational attainment in the United States: 2015, US census bureau report, US Census Bureau.

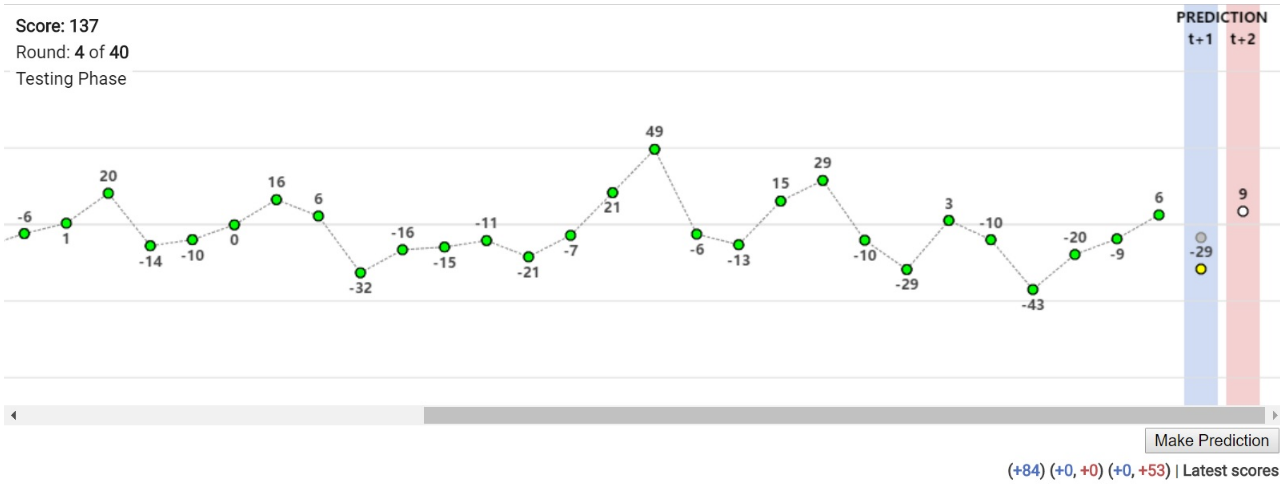
Shiller, Robert, 1981, Do stock prices move too much to be justified by subsequent changes in dividends?, *American Economic Review* .

Tversky, Amos, and Daniel Kahneman, 1974, Judgment under uncertainty: Heuristics and biases, *Science* 185, 1124–1131.

Woodford, Michael, 2003, Imperfect common knowledge and the effects of monetary policy, *Knowledge, Information, and Expectations in Modern Macroeconomics* .

Figures

Figure 1: Prediction Screen



Score: 137

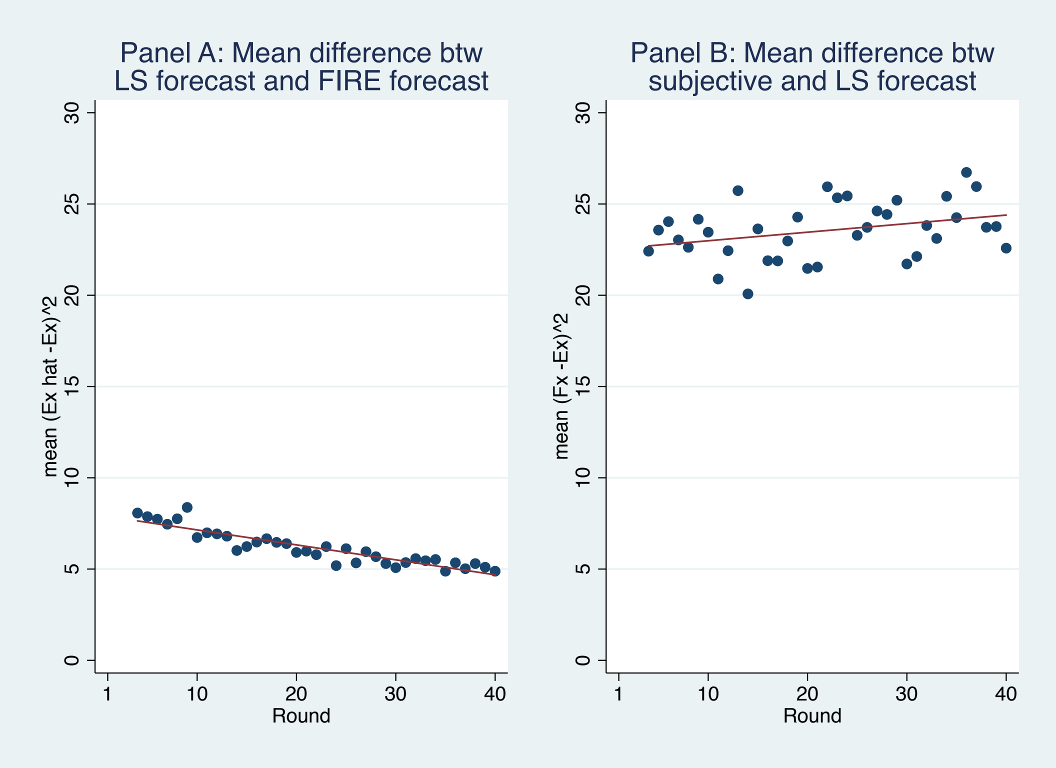
Round: 4 of 40

Testing Phase

(+84) (+0, +0) (+0, +53) | Latest scores

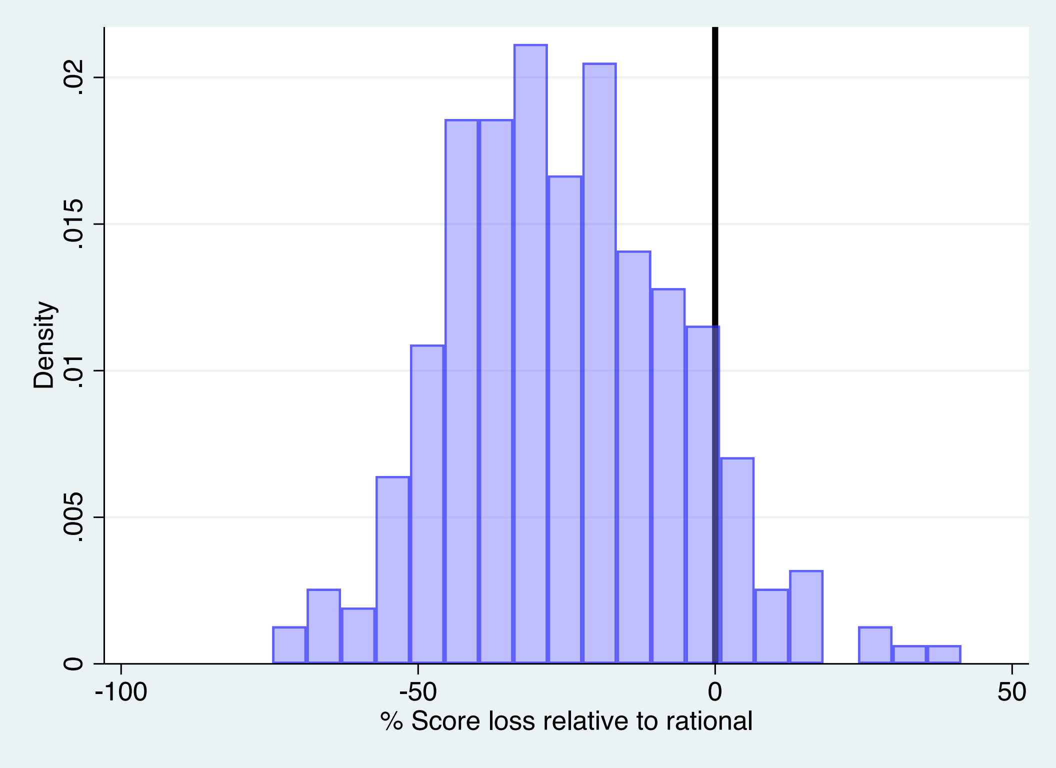
Note: Screen shot of the prediction task. The green dots indicate past realizations of the statistical process. In each round t, participants are asked to make predictions about two future realizations $F_t x_{t+1}$ and $F_t x_{t+2}$. They can drag the mouse to indicate $F_t x_{t+1}$ in the purple bar 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 2: Distance between Subjective Forecasts and Rational Expectations



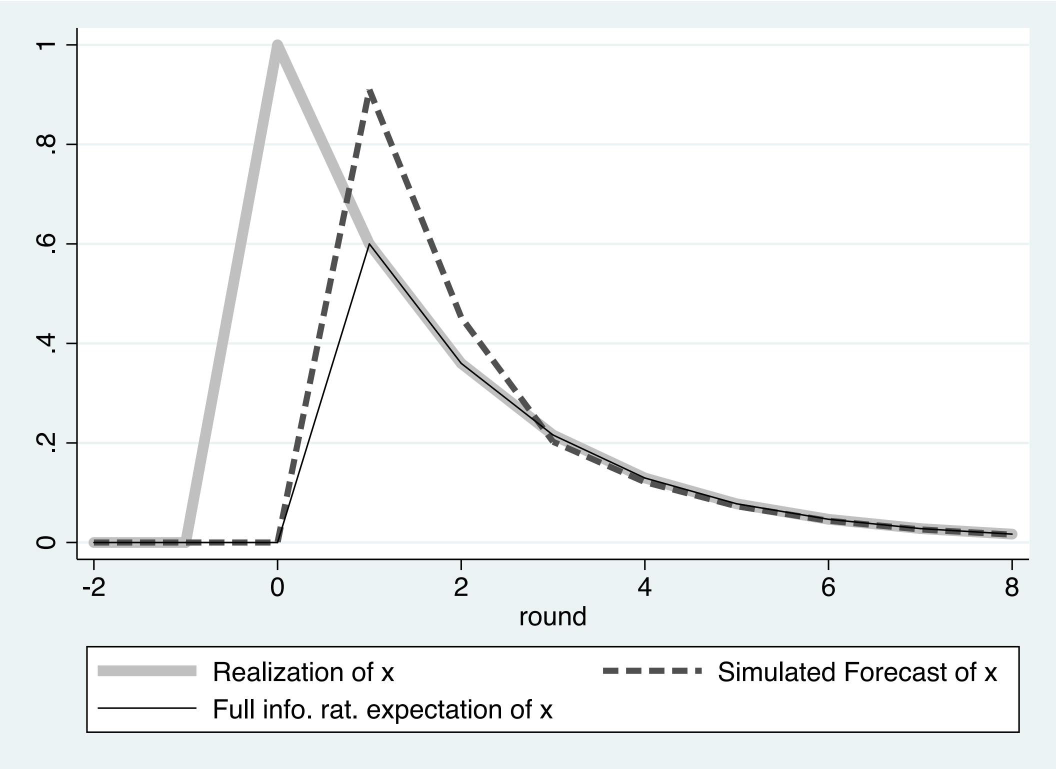
Note: Panel A shows the root mean squared difference between least squares (LS) expectations and full information rational forecasts (FIRE). Panel B shows the root mean squared difference between participants' actual subjective forecasts and LS rational forecasts. The data use all conditions in Experiment 1. For each round t from 1 to 40, we compute the mean squared difference between the subjective forecast $F_{t-1}x_t$ and the full information rational forecast $E_{t-1}x_t = \rho x_{t-1}$. We then take the square root of this, and report it in Panel A. Hence, in Panel A, if all survey participants were full information rational, the mean difference would be equal to zero. We repeat this procedure in Panel B, replacing the subjective forecast with the LS learning expectation $E^{LS}_{t-1}x_t$ obtained by regressing x_s on x_{s-1} for all periods between -40 and $t-1$. Hence, Panel B shows the extent to which a LS learner would converge to the full information rational expectation. The root mean squared difference between FI RE and LS learning forecasts goes down from 7 to 5 after 40 rounds. The root mean squared difference between subjective forecasts and LS forecasts goes up from 22 to 24 after 40 rounds.

Figure 3: Participants' Score Losses compared to LS Rational Expectations



Note: This Figure shows the histogram of participants' scores, benchmarked by the score that a LS learner would have obtained in the same data. The sample here is Experiment 1. For each participant in Experiment 1, in each period, we estimate LS rational expectations using 3 lags and all past observations. We then construct the score S^* with such LS learning. We report here the distribution of the percentage score loss of subjective expectations, computed as $(S - S^*)/S^*$ where S is the actual score.

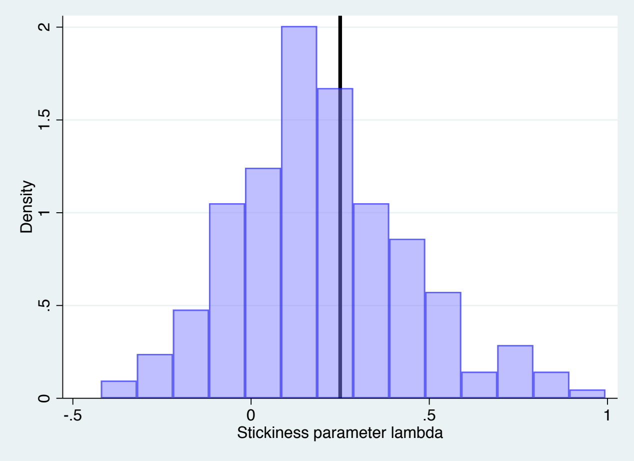
Figure 4: Expectation Response to an Impulse in x



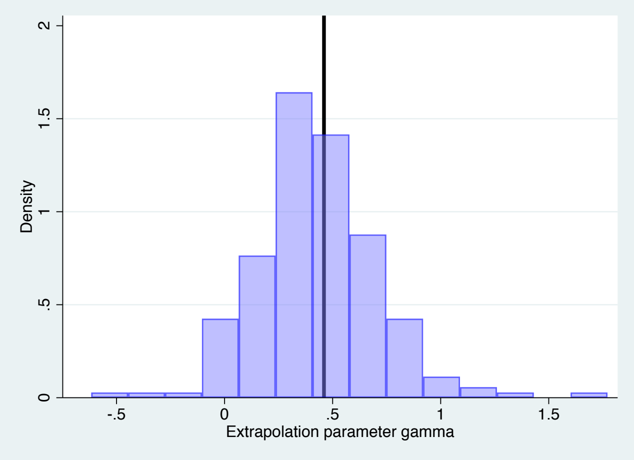
Note: For the case of $\rho = .6$, we show the impulse response of the process x, the full information rational expectations, and forecasts implied by the formulation estimated in Equation (7). The thick light grey line corresponds to the simulation of the response of an AR(1) x_t to a one time shock in ϵ equal to 1. Hence, $x_0 = 1$ and for each $t \geq 1$, $x_t = .6x_{t-1}$. The fine dark line is the full information rational expectation, equal to $E_{t-1}x_t = 0$ until $t = 0$, and equal to $E_{t-1}x_t = \rho x_{t-1}$ for $t \geq 1$. The dark dashed line corresponds to the forecasting process estimated in the paper. To parametrize expectations formation, we use $\lambda = .25$ and $\gamma = .46$ as suggested by the results from Table 5.

Figure 5: Distribution of Stickiness and Extrapolation Estimates across Individuals

Panel A: Distribution of Stickiness λ



Panel B: Distribution of Extrapolation γ



Note: For participants in Experiment 1 (Baseline, MTurk), we run the following regression individual by individual:

$$F_t^i x_{it+1} - E_t x_{it+1} = \lambda_i \left(F_{t-1}^i x_{it+1} - E_t x_{it+1} \right) + \gamma_i (x_{it} - E_{t-1} x_{it}) + u_{it+1}$$

where λ_i and γ_i are allowed to differ across subjects. $E_t x_{it+1}$ is computed using the LS learning definition. We then report the distribution of these parameters in the two panels above. The vertical black line corresponds to the estimates of the average model in Table 5, column (3).

Tables

Table 1: Experimental Literature on Expectations Formation

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Paper	# of participants	# of history	# of predictions	Process	Same draw	Econ. background	Monetary Incentives	Format graph /text	Forecast Horizon	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
Beshears et al(2013)	98	100k	60	ARIMA (0,1,50)	No	No	Yes	Graph	1	Natural Expec.
Frydman &Nave(2016)	38	10	400	$\rho \approx 1$	?	Yes	Yes	Graph	1	Extrap.
This paper	1,500+	40	40	$\rho \in \{0, .2, .4, .6, .8, 1\}$	Both	Both	Yes	Both	1,2,5	Sticky+ Extrapolation.

Note: 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. Nearly all experiments feature some form of monetary incentives (column (8)). Column (9) describes the format used to present the data (graphical or number list). Column (10) shows the forecast horizon requested (most of the time, just a one period ahead forecast). The last column describes the models tested: Most of the time, either adaptive expectation or some form of extrapolation.

Table 2: Summary of Conditions in All Experiments

		(1)	(2)	(3)	(4)	(5)	(6)
#	Short		AR(1) process		Forecasts	Grey dot	Number
	description	persistence	constant	volatility	asked		of
		ρ	μ	σ_ϵ			participants
Panel A: Experiment 1 – Baseline, MTurk							
A1	Baseline $\rho = 0$	0	0	20	F1+F2	Y	32
A2	Baseline $\rho = .2$	.2	0	20	F1+F2	Y	32
A3	Baseline $\rho = 0.4$	.4	0	20	F1+F2	Y	36
A4	Baseline $\rho = 0.6$	.6	0	20	F1+F2	Y	39
A5	Baseline $\rho = 0.8$	.8	0	20	F1+F2	Y	28
A6	Baseline $\rho = 1$	1	0	20	F1+F2	Y	40
Panel B: Experiment 2 – Common path, MTurk							
B1	Path 1	.6	0	20	F1+F2	Y	37
B2	Path 2	.6	0	20	F1+F2	Y	32
B3	Path 3	.6	0	20	F1+F2	Y	37
B4	Path 4	.6	0	20	F1+F2	Y	30
B5	Path 5	.6	0	20	F1+F2	Y	32
B6	Path 6	.6	0	20	F1+F2	Y	33
B7	Path 7	.6	0	20	F1+F2	Y	27
B8	Path 8	.6	0	20	F1+F2	Y	33
B9	Path 9	.6	0	20	F1+F2	Y	26
B10	Path 10	.6	0	20	F1+F2	Y	43
Panel C: Experiment 3 – Robustness checks, MTurk							
C1	Context: quarterly GDP growth	.4	.40	.55	F1+F2	Y	38
C2	Context: monthly inflation	.4	.12	.23	F1+F2	Y	39
C3	Context: monthly stock returns	.2	.55	3.43	F1+F2	Y	29
C4	Context: monthly house price growth	.8	.02	.39	F1+F2	Y	37
C5	No context, comparison	.4	.40	.55	F1+F2	Y	30
C6	No context, comparison	.4	.12	.23	F1+F2	Y	34
C7	No context, comparison	.2	.55	3.43	F1+F2	Y	36
C8	No context, comparison	.8	.02	.39	F1+F2	Y	35
C9	Comparison	.4	0	20	F1+F2	Y	30
C10	Change horizon	.2	0	20	F1	/	37
C11	Change horizon	.4	0	20	F1	/	36
C12	Change horizon	.6	0	20	F1	/	33
C13	Change horizon	.8	0	20	F1	/	38
C14	Change horizon	.2	0	20	F2	Y	38
C15	Change horizon	.4	0	20	F2	Y	51
C16	Change horizon	.6	0	20	F2	Y	32
C17	Change horizon	.8	0	20	F2	Y	42
C18	Change horizon	.2	0	20	F1+F5	Y	27
C19	Change horizon	.4	0	20	F1+F5	Y	34
C20	Change horizon	.6	0	20	F1+F5	Y	29
C21	Change horizon	.8	0	20	F1+F5	Y	41
C22	No grey dot	.2	0	20	F1+F2	N	26
C23	No grey dot	.4	0	20	F1+F2	N	31
C24	No grey dot	.6	0	20	F1+F2	N	30
C25	No grey dot	.8	0	20	F1+F2	N	42
Panel D: Experiment 4 – DGP information, MIT EECS							
D1	Baseline	.2	0	20	F1+F2	Y	42
D2	Baseline	.6	0	20	F1+F2	Y	52
D3	Display DGP is AR(1)	.2	0	20	F1+F2	Y	70
D4	Display DGP is AR(1)	.6	0	20	F1+F2	Y	40

Note: This Table provides a summary of the 4 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 $x_{t+1} = \mu + \rho x_t + \epsilon_{t+1}$. Column (4) shows the forecasts asked to each participants. For example, "F1+F2" means one- and two-period ahead forecasts. Column (5) indicates if a grey dot is present on the interface to indicate the long-term forecast participants made previously. Column (6) reports the number of participants. Typically, each participant is presented with a different draw, except in Experiment 2, where all participants within a given condition are presented with the same draw. Participants are only allowed to participate once.

Table 3: Experimental Statistics

	Mean	p25	p50	p75	SD	*N*
Experiment 1 (2 forecasts per round)						
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
per round (sec)	10.64	6.74	8.66	11.77	6.59	270
Total forecast score	1,996	1,680	1,985	2,305	469	270
Bonus ($)	3.33	2.80	3.31	3.84	.78	270
Experiment 2 (2 forecasts per round)						
Total time (min)	13.12	8.16	10.88	15.79	7.88	330
Forecast time (min)	6.78	4.59	5.75	7.74	4.05	330
per round (sec)	10.17	6.89	8.63	11.61	6.07	330
Total forecast score	1,932	1,645	1,890	2,205	502	330
Bonus ($)	3.22	2.74	3.15	3.67	.84	330
Experiment 3 (2 forecasts per round)						
Total time (min)	12.44	7.83	10.56	14.60	7.48	580
Forecast time (min)	6.69	4.42	5.73	7.67	4.17	580
per round (sec)	10.03	6.63	8.59	11.50	6.26	580
Total forecast score	1,973	1,667	1,971	2,276	484	580
Bonus ($)	3.29	2.78	3.29	3.80	.81	580
Experiment 3 (1 forecast per round)						
Total time (min)	11.05	6.91	9.25	13.73	6.52	295
Forecast time (min)	5.27	3.36	4.31	6.03	3.80	295
per round (sec)	7.91	5.03	6.47	9.05	5.70	295
Total forecast score	974	789	974	1,149	289	295
Bonus ($)	1.62	1.31	1.62	1.92	.48	295
Experiment 4 (2 forecasts per round)						
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
per round (sec)	13.17	6.05	7.64	11.19	29.58	204
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

Note: This Table reports basic experimental statistics for each of the 4 experiments we conducted. Experiment 3 has some conditions with 1 forecast per round and some conditions with 2 forecasts per found, as shown in Table 2. We report the distributions of the following statistics: the overall time taken to complete the experiment, the time taken to complete the forecasting part (the main part), the total score, and the total bonus paid in US dollars.

Table 4: Demographics

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
		Experiment 1		Experiment 2		Experiment 3		Experiment 4	
		Obs.	%	Obs.	%	Obs.	%	Obs.	%
Panel A: Demographic Characteristics									
Gender	Male	151	55.9	201	60.9	457	52.2	88	43.1
	Female	119	44.1	129	39.1	418	47.8	116	56.9
Age	<= 25	36	13.3	44	13.3	129	14.7	197	96.6
	25-45	186	68.9	224	67.9	593	67.8	7	3.4
	45-65	44	16.3	57	17.3	145	16.6	0	0
	65+	4	1.5	5	1.5	8	.9	0	0
Education	Grad school	26	9.6	42	12.7	121	13.8	0	0
	College	170	63.0	200	60.6	524	59.9	207	100.0
	High school	74	27.4	88	26.7	224	25.6	0	0
	Below/other	0	.0	0	.0	6	.7	0	0
Invest. exper.	Extensive	7	2.6	6	1.8	23	2.6	2	1
	Some	71	26.3	74	22.4	193	22.1	43	21.1
	Limited	100	37.0	129	39.1	367	41.9	138	67.7
	None	92	34.1	121	36.7	292	33.4	21	10.3
Taken stat class	Yes	110	40.7	144	43.6	406	46.4	-	-
	No	160	59.3	186	56.4	469	53.6	-	-
Panel B: Statistical Tests									
Median test	Correct	138	51.1	167	50.6	426	48.7	182	89.2
	Incorrect	132	48.9	163	49.4	449	51.3	22	10.8
Hospital test	Larger hospital	91	33.7	98	29.7	308	35.2	18	8.8
	Smaller hospital (correct)	92	34.1	118	35.8	288	32.9	119	58.3
	Same	87	32.2	114	34.6	279	31.9	67	32.8
Coin toss test	Mix more likely	85	31.5	75	22.7	221	25.3	10	4.9
	Trend more likely	16	5.9	15	4.6	54	6.2	4	1.7
	Same (correct)	198	62.2	587	67.1	587	67.1	187	91.7
	None	1	.4	2	.6	13	1.5	3	1.5
Total		270	100.0	330	100.0	875	100.0	204	100.0

Note: This Table describes demographics of participants in Panel A, and answers to test questions on statistics in Panel B. Columns (1) and (2) provide information for the participants to Experiment 1 (Baseline, MTurk); columns (3) and (4) for Experiment 2 (Common path, MTurk); columns (5) and (6) for Experiment 3 (Robustness checks, MTurk); and columns (7) and (8) for Experiment 4 (Describe DGP, MIT EECS). "Median test" asks participants to enter the median of the following numbers: 10, 30, 60, 70, 90, 150, 220, 760. "Hospital test" asks participants the following question: "A town has two hospitals. The larger hospital has on average 35 babies born every day. The smaller hospital has on average 10 babies born every day. We know that about 50 percent of babies are boys. For a period of 6 months, the hospitals recorded the number of days when more than 70 percent of the babies born are boys, and called them 1baby boy days.' Which of the following do you think is most likely?" "Coin toss test" asks participants the following question: "A fair coin is tossed 6 times. What do you think about the likelihood of seeing Pattern A: H-T-H-T-T-H vs. Pattern B: H-H-H-T-T-T?"

Table 5: Expectation Formation Model: Main results

	$F_tx_{t+1} - \widehat{E}_tx_{t+1}$						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Sticky	Extrap	Main		$FIRE$	$t \leq 20$	$t > 20$
$F_{t-1}x_{t+1} - \widehat{E}_tx_{t+1}$	.16***		.25***		.24***	.24***	.25***
	(3.7)		(6)		(5.6)	(7.5)	(4)
$x_t - \widehat{E}_{t-1}x_t$		.36***	.46***	.46***	.44***	.43***	.48***
		(17)	(19)	(20)	(16)	(15)	(13)
$F_{t-1}x_{t+1}$				.24***			
				(5.9)			
$\widehat{E}_tx_{t+1}$				-.26***			
				(-6.6)			
N	8073	8280	8073	8073	8073	3726	4347
r2	.04	.099	.19	.19	.16	.18	.2

Note: We pool together all forecasts in Experiment 1 (Baseline, MTurk) for which $\rho \in \{0, .2, .4, .6, .8, 1\}$, and run the main specification:

$$F_t^i x_{it+1} - E_t x_{it+1} = \lambda \left(F_{t-1}^i x_{it+1} - E_t x_{it+1}\right) + \gamma(x_{it} - E_{t-1}x_{it}) + u_{it+1}$$

In all columns but column (5), we use Least Square RE $\widehat{E}_tx_{t+1}$ to measure rational expectations (in both dependent and independent variables). LS expectations are obtained using a rolling OLS regression of x_t on three lags, separately for each individual. In column (1), we set $\gamma = 0$. In column (2), we set $\lambda = 0$. Column (3) is our main specification. Column (4) allows γ to differ for the two components of the sticky term ($F_{t-1}x_{t+1}$ and E_tx_{t+1}). Column (5) uses full information rational expectations instead of LS expectations. FIRE is $E_tx_{it+1} = \rho x_{it}$ using the true ρ. Columns (6) and (7) split the sample into the first and last 20 rounds. t-stats in parentheses; standard errors clustered by individual.

Table 6: Explaining Average Expectations

	$F_t x_{t+1} - \widehat{E}_t x_{t+1}$						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Sticky	Extrap	Main		FIRE	$t \leq 20$	$t > 20$
$F_{t-1} x_{t+1} - \widehat{E}_t x_{t+1}$	-.11***		.29***		.21***	.32***	.26***
	(-4.1)		(12)		(7.2)	(8.9)	(8.1)
$x_t - \widehat{E}_{t-1} x_t$		.28***	.47***	.5***	.46***	.47***	.47***
		(14)	(21)	(16)	(16)	(13)	(17)
$F_{t-1} x_{t+1}$				.3***			
				(12)			
$\widehat{E}_t x_{t+1}$				-.34***			
				(-6.8)			
N	390	400	390	390	390	180	210
r2	.045	.4	.57	.58	.56	.53	.62

Note: This Table follows the structure of Table 5, except that the panel data now consists of the average forecasts from 10 given paths over 40 rounds. We use data from Experiment 2 (Common path, MTurk). For each path, the AR(1) process has $\rho = 0.6$ and around 33 participants make forecasts along each path. We average forecasts and expectations across participants facing the same path. We then run the following regression:

$$\bar{F}_t x_{ct+1} - E_t x_{ct+1} = \lambda \left(\bar{F}_{t-1} x_{ct+1} - E_t x_{ct+1} \right) + \gamma (x_{ct} - E_{t-1} x_{ct}) + u_{ct+1}$$

for condition c at round t. $\bar{F}_t x_{ct+1}$ is the *average* prediction across participants in condition c at round t for next period realization. In all columns but (5), rational expectations are computed using LS learning as we discuss in the paper.

Table 7: Main Expectation Formation Model:
Sample Splits by Participant Groups

Dependent variable	$F_t x_{t+1} - \widehat{E}_t x_{t+1}$					
Panel A : Socio-demographics						
	(1)	(2)	(3)	(4)	(5)	(6)
	Male	Female	Age< 35	Age≥ 35	High School	College
$F_{t-1}x_{t+1} - \widehat{E_t}x_{t+1}$	.27***	.21***	.25***	.24***	.23***	.25***
	(4.3)	(6.5)	(4.1)	(7.9)	(5.3)	(5)
$x_t - \widehat{E_{t-1}}x_t$	.45***	.45***	.46***	.45***	.47***	.45***
	(13)	(17)	(13)	(18)	(16)	(16)
N	4563	3510	4680	3393	2145	5928
r2	.19	.19	.17	.22	.21	.19
Test of equality – p value						
Stickiness	0.44		0.97		0.69	
Extrapolation	1.00		0.98		0.56	
Panel B : Answers to statistics quiz						
	(1)	(2)	(3)	(4)	(5)	(6)
	Coin toss		Median		Hospital	
	False	Right	False	Right	False	Right
$F_{t-1}x_{t+1} - \widehat{E_t}x_{t+1}$	.21***	.26***	.21***	.28***	.2***	.32***
	(6.3)	(4.5)	(6.1)	(3.9)	(9.4)	(3.5)
$x_t - \widehat{E_{t-1}}x_t$	.45***	.45***	.44***	.47***	.46***	.45***
	(13)	(15)	(16)	(13)	(21)	(8.3)
N	2925	5148	3939	4134	5304	2769
r2	.19	.19	.16	.22	.19	.21
Test of equality – p value						
Stickiness	0.43		0.38		0.19	
Extrapolation	0.94		0.44		0.95	

Note: This Table estimates our main specification in subsamples of our experiment, using data from Experiment 1 (Baseline, MTurk). In all columns, rational expectations are computed using the LS learning assumption we discuss in the paper (hence the "hat") in the expectation operator. We pool participants across $\rho \in \{0, .2, .4, .6, .8, 1\}$ and run the following regression:

$$F_t^i x_{it+1} - E_t x_{it+1} = \lambda \left(F_{t-1}^i x_{it+1} - E_t^i x_{it+1}\right) + \gamma(x_{it} - E_{t-1}^i x_{it}) + u_{it+1}$$

In Panel A, we focus on socio-demographic categories. In columns (1) and (2), we split the sample into male and female participants. In columns (3) and (4), we split the sample into participants above and below 35 years old. In columns (5) and (6), we split the sample into participants with high school degree and below and participants with college degree and above. In Panel B, we focus on groups by answers to various statistical questions. Columns (1) and (2) split participants into false and right answers to the "coin toss" question, which tests understanding of statistical independence. Columns (3) and (4) split participants into false and right answers to the "median" question, which tests is people can correctly compute the median of nine numbers. Columns 5 and 6 split participants into false and right answers to the "hospital" question, which tests the understanding of the law of large numbers. t-stats in parentheses; standard errors clustered by individual.

Table 8: Main Expectation Formation Model
Sample Split by Value of ρ

			$F_t x_{t+1} - \widehat{E}_t x_{t+1}$				Equality
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$\rho =$	0	.2	.4	.6	.8	1	Test
							p-value
Panel A : Baseline Model							
$F_{t-1}x_{t+1} - \widehat{E}_t x_{t+1}$	.17**	.16***	.12**	.24***	.21***	.42***	.14
	(2.5)	(3.5)	(2.3)	(6.1)	(6.4)	(3.9)	
$x_t - \widehat{E_{t-1}} x_t$	.44***	.46***	.48***	.42***	.41***	.58***	.53
	(8.9)	(7.8)	(13)	(9.9)	(9.8)	(6.4)	
N	1248	1248	1404	1521	1092	1560	
R^2	.19	.2	.16	.16	.15	.32	
Panel B : Extrapolation only							
$x_t - \widehat{E_{t-1}} x_t$	.45***	.44***	.44***	.31***	.27***	.25***	.00
	(8.8)	(7.7)	(12)	(5.3)	(5.6)	(5.1)	
N	1280	1280	1440	1560	1120	1600	
R^2	.15	.17	.15	.067	.084	.045	
Panel C : Stickiness only							
$F_{t-1}x_{t+1} - \widehat{E}_t x_{t+1}$	.18**	.13***	.038	.16***	.062	.27**	.09
	(2.7)	(2.8)	(.89)	(4.3)	(1.6)	(2.2)	
N	1248	1248	1404	1521	1092	1560	
R^2	.039	.022	.0019	.048	.0087	.14	
Panel D : Extrapolation & Stickiness, backward-looking (replacing $\widehat{E}_t x_{t+1}$ by x_t)							
$F_{t-1}x_{t+1} - x_t$	.24***	.23***	.15**	.24***	.22***	.44***	.41
	(3.4)	(5.3)	(2.4)	(5)	(7.1)	(3.8)	
$x_t - x_{t-1}$	-.13**	.0089	.071	.18***	.29***	.58***	.00
	(-2.7)	(.17)	(1.4)	(4.9)	(6.8)	(5)	
N	1248	1248	1404	1521	1092	1560	
R^2	.18	.12	.033	.095	.094	.29	

Note: We estimate our model for each value of $\rho \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$, using data from Experiment 1 (Baseline, MTurk). In panels A,B,C, rational expectations are computed using the LS learning model (hence the "hat" above the expectation operator). Panel A reports the result for the main model. Panels B and C only use one of the indepedent variables. Panel D uses backward-looking expectation, i.e. replaces $E_t x_{t+1}$ with x_t. Column (1)-(5) show estimates for each value of ρ; column (6) reports the p-value of an equality test of all coefficients. t-stats in parentheses; standard errors clustered by individual.

Table 9: Sensitivity to Process Parameters beyond Changes in ρ

Dependent variable	$F_t x_{t+1} - \widehat{E}_t x_{t+1}$				
Treatment	C9 (Main)	C5	C6	C7	C8
Constant $\mu =$	0	0.40	0.12	0.55	0.02
Persistence $\rho =$	0.4	0.4	0.4	0.2	0.8
Innovation vol. $\sigma =$	20	0.55	0.23	3.43	.39
	(1)	(2)	(3)	(4)	(5)
$F_{t-1}x_{t+1} - \widehat{E}_t x_{t+1}$	.24***	.22***	.29***	.34***	.39***
	(4.3)	(4.1)	(5)	(4.7)	(2.9)
$x_t - \widehat{E_{t-1}} x_t$	.48***	.39***	.38***	.51***	.53***
	(10)	(6.5)	(10)	(9.4)	(5.4)
N	1638	1170	1326	1404	1365
r2	.24	.18	.19	.25	.25
Test of equality with main setting – p value					
Stickiness	.	0.81	0.54	0.25	0.31
Extrapolation	.	0.28	0.13	0.62	0.61

Note: We test here whether changes in σ, the volatility of innovation, and μ, the constant of the process, affect our estimates of λ and γ. We use data from Experiment 3 (Robustness checks, MTurk). In each of these conditions, the process for participant i is given by $x_{it+1} = \mu + \rho x_{it} + \sigma \epsilon_{it}$ where ϵ_{it} is standardized normal. Each column corresponds to a condition where σ and μ are different. For each condition, we then run the regression:

$$F_t^i x_{it+1} - E_t x_{it+1} = \lambda \left(F_{t-1}^i x_{it+1} - E_t^i x_{it+1} \right) + \gamma (x_{it} - E_{t-1}^i x_{it}) + u_{it+1}$$

We then report, in the bottom panel, p-value of tests of equality between coefficients of these processes and the benchmark setting of column (1). These tests are done by running the two regressions as SURE. Each subject has a different draw of the process, so we cluster standard error by individual. In all columns, rational expectations are computed using the LS learning assumption we discuss in the paper. t-stats in parentheses.

Table 10: Sensitivity to Term Structure Reporting

Dependent variable	$F_t^i x_{it+1} - \widehat{E}_t x_{it+1}$		$F_t x_{t+1}$			$F_{t-1} x_{t+1}$	
	Grey dot		Effect of reporting LT expec.			Effect of reporting ST expec.	
	Main setting	Without	Main setting	$F_t x_{t+1}$ only	$F_t x_{t+1}$ & $F_t x_{t+5}$	Main setting	$F_t x_{t+2}$ only
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$F_{t-1} x_{t+1} - \widehat{E_t x_{t+1}}$	.24***	.081					
	(4.3)	(1.7)					
$x_t - \widehat{E_{t-1}} x_t$	.48***	.45***					
	(10)	(7)					
$\widehat{E_t x_{t+1}}$			.75***	.71***	.68***		
			(8.5)	(8.7)	(7.2)		
$\widehat{E_{t-1} x_{t+1}}$			.26***	.16*	-.0013	.82***	.78***
			(3)	(1.8)	(-.016)	(8.4)	(8.7)
$\widehat{E_{t-2} x_{t+1}}$			.0077	.085	.18***	.073	.09
			(.14)	(1.4)	(2.9)	(.79)	(1)
$x_t - \widehat{E_{t-1} x_t}$			.54***	.41***	.43***		
			(13)	(8.8)	(8.4)		
$x_{t-1} - \widehat{E_{t-2} x_{t-1}}$			.0034	.021	.076***	.48***	.5***
			(.13)	(.92)	(3.1)	(13)	(15)
$x_{t-2} - \widehat{E_{t-3} x_{t-2}}$			-.0087	.037*	.00034	.067***	.092***
			(-.45)	(1.9)	(.018)	(2.7)	(3.4)
N	1638	1053	5016	5168	4864	5016	6042
r2	.24	.14	.49	.43	.39	.33	.27
Test of equality with main setting – p value							
Stickiness	.	0.03	.	0.70	0.59	.	0.75
Extrapolation	.	0.80	.	0.04	0.11	.	0.68

Note: next page

Sensitivity to Term Structure Reporting (Cont'd)

Note: We test the impact of reporting different term structures of expectations using data from Experiment 3 (Robustness checks, MTurk). In columns (1) and (2), we test whether the presence of a gray dot, to help subjects remember their previous two-period-ahead forecast, $F_{t-1}x_{t+1}$, affects expectation formation. In column (1), we report the results of our baseline setting, identical to Table 9, column (1). In column (2), we use exactly the same parameters and setting but remove the grey dot which reminds participants of $F_{t-1}x_{t+1}$ when they report $F_t x_{t+1}$ and $F_t x_{t+2}$. We run our main specification (7) for both conditions and test the equality of coefficients in the bottom panel. In columns (3)-(5), we test whether the reporting of long-term expectations affects the reporting of short-term ones. In these columns, we use the specification with lags in (8), where we regress $F_t x_{t+1}$ on lagged values of rational expectations $E_{t-k}x_{t+1}$ and past innovations of $x_{t-k} - E_{t-k-1}x_{t-k}$ for $k \geq 0$. Under our recursive model (7), the coefficient on the first lags $E_t x_{t+1}$ and $x_t - E_{t-1}x_t$ are equal to $1 - \lambda$ and γ respectively. In column (3), we report the baseline condition of column (1), but using the "lag" regression specification. In column (4), we report the condition where participants make forecasts about the same process, but are not required to report the long-term expectation $F_t x_{t+2}$. In column (5), on the contrary, participants are asked to report very long-term expectations $F_t x_{t+5}$. We test equality of these coefficients with estimates of column (3) in the bottom panel. In columns (6)-(7), we test the effect of reporting short-term expectations on long-term ones. The methodology is identical to columns (3)-(5), except that now we regress $F_{t-1}x_{t+1}$ on lagged values of rational expectations $E_{t-k}x_{t+1}$ and past innovations of $x_{t-k} - E_{t-k-1}x_{t-k}$ for $k \geq 1$. The coefficients on $E_{t-1}x_{t+1}$ and $x_{t-1} - E_{t-2}x_{t-1}$ are in theory equal to $1 - \lambda$ and γ. In column (6), we report regression results for the baseline condition as in columns (1) and (3). In column (7), we report regression results for the condition where participants are only required to forecast $F_t x_{t+1}$ – thus not the short-term expectation. We test equality of coefficients on the first lags in the bottom panel. Each participant has a different draw of the process, so we cluster standard errors by individual. In all columns, rational expectations are computed using the LS learning assumption we discuss in the paper. t-stats in parentheses.

Table 11: Sensitivity to Context

	$F_t x_{t+1} - \widehat{E}_t x_{t+1}$							
	"GDP growth"		"Inflation"		"Stock returns"		"House price"	
	Without	With	Without	With	Without	With	Without	With
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$F_{t-1}x_{t+1} - \widehat{E_t x_{t+1}}$	.22***	.29***	.29***	.31***	.34***	.21**	.39***	.26***
	(4.1)	(8.3)	(5)	(6.5)	(4.7)	(2.6)	(2.9)	(3.2)
$x_t - \widehat{E_{t-1}x_t}$	.39***	.47***	.38***	.37***	.51***	.44***	.53***	.45***
	(6.5)	(9.1)	(10)	(7.3)	(9.4)	(6.7)	(5.4)	(5.1)
N	1170	1482	1326	1521	1404	1131	1365	1443
r2	.18	.21	.19	.21	.25	.21	.25	.17
Test of equality – p value								
Stickiness	0.29		0.83		0.21		0.40	
Extrapolation	0.32		0.77		0.42		0.55	

Note: We test here whether forecast behavior is similar for abstract processes and variables with economic context. For each condition, we run the following regression:

$$F_t^i x_{it+1} - \widehat{E}_t x_{it+1} = \lambda \left(F^i_{t-1} x_{it+1} - \widehat{E}^i_t x_{it+1}\right) + \gamma(x_{it} - \widehat{E}^i_{t-1} x_{it}) + u_{it+1}$$

In columns (1)-(2) we study US quarterly real GDP growth. We estimate an AR(1) using quarterly real GDP growth data, which leads to $x_t = .40 + .4x_{t-1} + .55\epsilon_t$ where $\epsilon_t \sim N(0,1)$. We then simulate one path per individual using these parameters. In column (1), like in our main tests in Table 8, participants are told that the process is a random process. In column (2), we write in the instruction at the beginning that "The process you will see has the same property as quarterly US real GDP growth." We then report p-values of equality tests of λ and γ across conditions in the bottom panel of the Table. In columns (3)-(4), we simulate a process estimated on monthly US CPI inflation. In columns (5)-(6), we simulate a process estimated on monthly S&P 500 returns. In columns (7)-(8), we simulate a process estimated on monthly house price growth. Each participant has a different draw of the process, so we cluster error terms at the individual level. In all columns, like in the rest of the paper, rational expectations are computed using the LS learning assumption we discuss in the paper (hence the "hat" above the expectation operator). t-stats in parentheses.

Table 12: Knowledge of the Prior

	$F_t x_{t+1} - \widehat{E}_t x_{t+1}$			
	(1)	(2)	(3)	(4)
	All		Understands AR(1)	
	Control	Treatment	Control	Treatment
$F_{t-1}x_{t+1} - \widehat{E_t x_{t+1}}$	.23***	.23***	.2***	.24***
	(9.9)	(10)	(6.1)	(9)
$x_t - \widehat{E_{t-1} x_t}$	.47***	.5***	.45***	.51***
	(16)	(19)	(11)	(12)
N	3666	4290	1794	1755
r2	.25	.23	.25	.25
Stickiness	0.98		0.47	
Extrapolation	0.53		0.28	

Note: We test here if more information about the data generating process affects expectations formation, using data from Experiment 4 (MIT EECS). 200 participants are randomly assigned to AR(1) processes with $\rho = .2$ or $\rho = .6$ (same processes as in the baseline Experiment 1). In addition, half of the participants are told that the process is an AR(1) process while the other half are not (they are told the process is a random process as in the baseline Experiment 1). In columns (1) and (2), we use the entire sample of participants. In columns (3) and (4), we focus on participants who perform well in answering test questions about AR(1) processes at the end of the experiment. t-stats in parentheses; standard errors clustered by individual.

APPENDIX – FOR ONLINE PUBLICATION

A Additional tests

A.1 Regime-Switching Beliefs

In this Appendix, we explore the possibility that participants have priors that x_t follows a regime-switching model (Barberis et al., 1998; Rabin, 2002; Rabin and Vayanos, 2010). In such a model, over-reaction can happen when agents observe a series of innovations with the same sign — which would lead them to believe that a trending regime has concurred. On the other hand, with only one innovation, such agents may appear to under-react, as one shock does not shift prior about regime very much. We implement here a test that follows this logic using data in Experiment 1 (Baseline, MTurk).

Before doing this, however, we note that the evidence from MIT EECS students does not point towards such prior. Indeed, Table 12 shows that knowing that the process is generated through a fixed, stationary AR(1) model does not affect expectations formation significantly (expectation formation has the same parameters λ and γ whether participants are told that the process is AR(1) or not).

The overall idea of the test is that, under the "regime-switching" priors, over-reaction should be stronger when there has been a sequence of innovations of the same sign. On the contrary, agents observing only one innovation in a direction would appear to under-react. We implement this in the context of our core specification (6) where:

$$F_t x_{t+1} - E_t x_{t+1} = \lambda \left(F_{t-1} x_{t+1} - E_t x_{t+1} \right) + \gamma (x_t - E_{t-1} x_t)$$

except that we allow the over-reaction to innovation γ to be higher when there has been a sequence of past innovations with the same sign. To to this, we split the innovation variable $\epsilon_t = x_t - E_{t-1} x_t$ into positive and negative realizations. We then allow the coefficient on positive realizations to be stronger if past realization ϵ_{t-1} is positive. Symmetrically for negative realizations. We can also go back one more period to $t-2$.

We report the results of this investigation in Table A.1. Column (1) is the same as the main specification in Table 5. Column (2) splits the innovation variable ϵ_t into positive and negative realization, and finds a symmetrical effect. Column (3) is the first test of regime-switching priors, by interacting realization with the sign of past realization. We find slight evidence that over-reaction is stronger with sequences of positive realizations, but the effect is asymmetric. Column (4) adds one more period but finds no evidence that over-reaction is stronger following three consecutive innovations with the same sign (as one would expect with regime-switching beliefs). In column (5), we remove the stickiness component. This makes the reaction to two consecutive positive realizations slightly bigger, but the estimated effect remains asymmetric. Also, the explanatory power of this model is much smaller (with an R^2 of .13 instead of .17), which suggest that the stickiness coefficient captures data variability better.

Overall, we note that there is some fragile evidence of regime-switching beliefs, but it is asymmetric and does not add significant explanatory power to our baseline model.

Table A.1: Allowing for Beliefs in Regime-Switch Models

	$F_t x_{t+1} - \widehat{E}_t x_{t+1}$				
	(1)	(2)	(3)	(4)	(5)
$F_{t-1}x_{t+1} - \widehat{E_t x_{t+1}}$	.19***	.18***	.18***	.18***	
	(8.2)	(8.1)	(7.6)	(7.5)	
$x_t - \widehat{E_{t-1}x_t} \equiv \epsilon_t$	.44***				
	(20)				
$\epsilon_t \times 1_{\epsilon_t \geq 0}$		.44***	.36***	.36***	.24***
		(13)	(9)	(9)	(6)
$\epsilon_t \times 1_{\epsilon_t \geq 0} \times 1_{\epsilon_{t-1} \geq 0}$			.15***	.15**	.24***
			(3.6)	(2.5)	(4.3)
$\epsilon_t \times 1_{\epsilon_t \geq 0} \times 1_{\epsilon_{t-1} \geq 0} \times 1_{\epsilon_{t-2} \geq 0}$				-.0012	.02
				(-.018)	(.31)
$\epsilon_t \times 1_{\epsilon_t \leq 0}$		.44***	.4***	.4***	.31***
		(13)	(11)	(11)	(8.1)
$\epsilon_t \times 1_{\epsilon_t \leq 0} \times 1_{\epsilon t-1 \leq 0}$			.065	.018	.089*
			(1.6)	(.37)	(1.8)
$\epsilon_t \times 1_{\epsilon_t \leq 0} \times 1_{\epsilon_{t-1} \leq 0} \times 1_{\epsilon_{t-2} \leq 0}$				.083	.11**
				(1.5)	(2)
N	6513	6346	6346	6346	6346
r2	.17	.17	.17	.17	.13

Note: In this Table, we experiment with various interactions of past changes in x_t in order to explore possible beliefs in regime-switching models. For each observation, we first construct the (full information) innovation on x_t as $x_t - \rho x_{t-1}$. We then ask whether expectation error reacts more to sequences of positive innovations, rather than single innovations.

A.2 Formulation with lags

In this appendix, we explore the alternative non-recursive specification of our empirical model of expectation formation.

First, note that our recursive specification in Equation (7) is equivalent to:

$$F_t x_{t+1} = (1-\lambda)\sum_{k\geq 0}\lambda^k E_{t-k}x_{t+1} + \gamma\sum_{k\geq 0}\lambda^k\left(x_{t-k-1} - E_{t-k-2}x_{t-k-1}\right) \tag{13}$$

Thus, the above relationship can be used to estimate λ and γ. We do this through non-linear least squares and report the results in Table A.2. We use data from Experiment 1 (Baseline, MTurk) below. Like in the rest of the paper, we allow for arbitrary correlation of error terms within individual. We experiment with different numbers of lags and separately estimate the parameters for the first 20 and last 20 periods of forecasts. The investigation leads to very stable parmeters for both γ and λ, mostly indistinguishable from our baseline estimates of Table 5.

Table A.2: Modeling Expectation Formation
Model with lags

	(1)	(2)	(3)	(4)	(5)
	4 lags	3 lags	2 lags	$t \leq 20$	$t > 20$
γ	0.45	0.45	0.42	0.43	0.46
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
λ	0.21	0.21	0.13	0.25	0.17
	(0.03)	(0.03)	(0.02)	(0.03)	(0.03)
R^2	0.11	0.11	0.11	0.12	0.11
Observations	7659	7659	7659	3312	4347

Note: On the panel of participants in Experiment 1 for which $\rho \in \{0, .2, .4, .6, .8, 1\}$, we estimate the following model:

$$F_t x_{t+1} = (1-\lambda)\sum_{k=0}^{n}\lambda^k E_{t-k}x_{t+1} + \gamma\sum_{k=0}^{n}\lambda^k\left(x_{t-k} - E_{t-k-1}x_{t-k}\right)$$

using non linear least squares. In columns (1), (2), (3), we assume respectively 4, 3, and 2 lags. In columns (4) and (5), we split the sample between the first and last 20 rounds of testing. Error terms are clustered at the individual level. Standard errors in parentheses.

B Survey Appendix

B.1 Sample Experiment

Below are the instructions for a sample experiment (Experiment 1, $\rho = .6$, $\mu = 0$, $\sigma = 20$). Participants first see a consent form with brief descriptions of the study. Once they agree to the consent, they will proceed to experimental instructions. The experiment starts with the forecasting task and is then followed by demographic questions. The demographic questions are the same for all of our experiments. The forecasting task may differ slightly depending on the treatment condition, as described in Section 2. We discuss these variants in the next subsection.

The experiment starts with a consent form:

Consent Form

Purpose of research: The purpose of this research is to study how people make predictions.
What you will do in this research: You will make forecasts about future realizations of a random process on a web-based platform, followed by a few demographics questions. There are 40 rounds, and you will make 2 predictions per round. You may exit the platform at any time or skip some questions without penalty.
Time required: It takes about 20 minutes to complete the study. You are free to spend as much time as you like up to 60 minutes.
Risks: There are no anticipated risks associated with participating in this study.
Compensation: You will receive **base payment** of **$1.80**. You will also receive a **bonus payment**. The **bonus payment** will be on the scale of **$2.50**, but the precise amount will depend on the accuracy of your predictions.

Your **base payment** and **bonus payment** will be distributed together within one week via MTurk.

Please feel free to contact us with the contact information below or through MTurk if you have any questions about payments. A summary of your payments will be displayed at the end of the study. You may save that page for your records.
Confidentiality: The system allows us to see MTurk Worker IDs and IP addresses. We may use these information for handling payments and to verify data quality, for example that you are in the United States and have not taken our previous surveys. Please make sure to mark your Amazon Profile as private if you do not want it to be found from your MTurk Worker ID. If you communicate with us via email to discuss any issue related to your participation, we will keep your information confidential. All personally identifiable information will be handled in compliance with Harvard and MIT data security requirements, will not be accessible to anyone outside the study team, and will not be used in our data analysis. Data analysis will be based on de-identified data. Part or all of the de-identified data may be shared with other researchers or be made available publicly for academic replication after publication.
Benefits: Your input will help our research develop a better understanding about how people make forecasts. We appreciate your participation. We hope you will also find the survey questions to be interesting.
Contact: If you have any questions, concerns, or suggestions related to this study, the researcher can be reached at:

David Thesmar Sloan School of Management, Massachusetts Institute of Technology 30 Memorial Dr, Cambridge, MA 02142 Cambridge, MA 02139 Email: thesmar@mit.edu (617) 324-7023

This research has been reviewed by the Committee on the Use of Human Subjects in Research at MIT. They can be reached at 617-253-6787, 77 Massachusetts Avenue, Room E25-143B, Cambridge, MA 02139, or couhes@mit.edu. You can contact them for any of the following:

- If your questions, concerns, or complaints are not being answered by the research team,
- If you cannot reach the research team,
- If you want to talk to someone besides the research team, or
- If you have questions about your rights as a research participant.

Please print or screenshot this page for your records.

By selecting to continue, you indicate that you are at least 18 years old and you agree to complete this HIT voluntarily.

[I Give My Consent]

(page break)

Experimental Instructions

Thank you very much for your participation. This study will take you about 20 minutes to complete.

You will receive base payment of **$1.80**. You will also receive a **bonus payment**. The typical bonus amount will be around **$2.50**, but the precise amount will depend on the accuracy of your predictions.

In this study, we would like to understand how people make predictions about future realizations of random processes. We will first show you 40 past realizations of a process, and you will make predictions of its future value for 40 rounds.

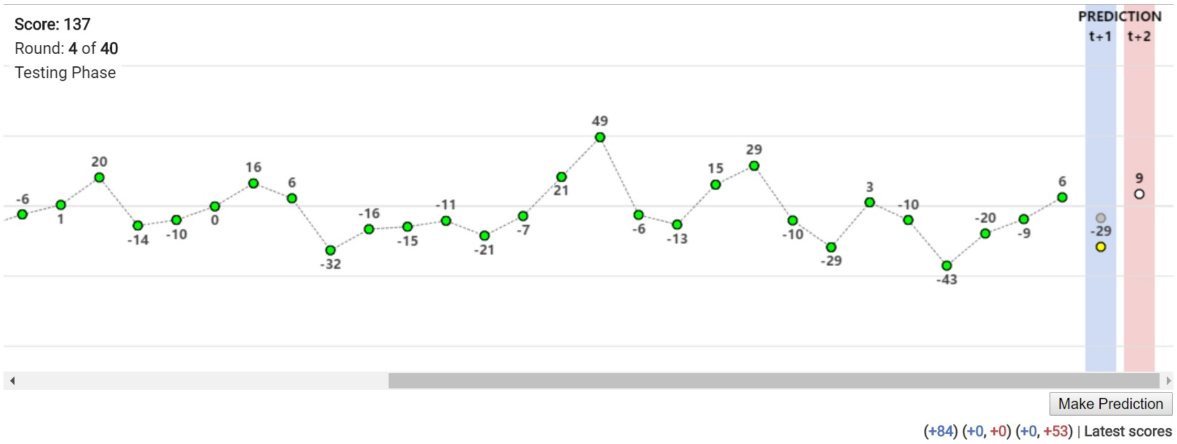
You will receive a score for each prediction you make. The more accurate your predictions are, the higher your score will be. If your prediction is out of a certain neighborhood around the actual value, you may receive a score of zero. The specific formula for the score of each prediction is $100 \times \max\{0, 1 - |\Delta|/20\}$ where Δ is the difference between your prediction and the realized value. We estimate that the best performer will receive an average score of 36 per prediction.

In each of the 40 rounds, we will ask you to predict the next two values of the process. At the end of the experiment, we will calculate your total score in the 40 rounds of predictions. *You will receive the bonus payment in U.S. dollars which is equal to your total score divided by 600.*

[Start Experiment]

(page break)

Experiment



Score: 137

Round: 4 of 40

Testing Phase

(+84) (+0, +0) (+0, +53) | Latest scores

(This plot is a screenshot of the interactive experimental interface. 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.)

Background Information

The prediction section is now over. We would now like to ask a few questions about yourself to help us in our research.

1. What is your gender?

- Male
- Female

2. What is your age?
3. What is the highest level of educational degree that you hold?
    - Graduate school (e.g. Masters, Ph.D., Post-doctoral degrees)
    - College
    - High school
    - Below high school
    - Other:
4. Have you taken statistics classes?
    - Yes
    - No
5. Do you have any experience investing in financial assets (e.g. stocks, bonds, mutual funds, pension funds, etc.)?
    - I have extensive experience investing in financial assets.
    - I have some experience.
    - I have very limited experience.
    - I have no experience at all.
6. What is the median of the following numbers? 10, 30, 60, 70, 90, 150, 220, 760
7. A town has two hospitals. The larger hospital has on average 35 babies born every day. The smaller hospital has on average 10 babies born every day. We know that about 50 percent of babies are boys. For a period of 6 months, the hospitals recorded the number of days when more than 70 percent of the babies born are boys, and called them “baby boy days." Which of the following do you think is most likely?
    - The larger hospital recorded more “baby boy days" than the smaller hospital.
    - The smaller hospital recorded more “baby boy days" than the larger hospital.
    - The two hospitals recorded the same number of “baby boy days."
8. A fair coin is tossed 6 times. What do you think about the likelihood of seeing Pattern A: H-T-H-T-T-H vs. Pattern B: H-H-H-T-T-T?
    - Pattern A is more likely than Pattern B
    - Pattern B is more likely than Pattern A
    - They are equally likely
    - None of the above
9. When would you say is a good time to invest in stocks:
    - If the stock market has been going up in the past two years
    - If the stock market has been going down in the past two years
    - I do not have an opinion

Feedback

The study is now completed. Do you have any comments and suggestions for the survey? Did you find anything to be unclear or confusing?

Submit Results

Click the button below to validate and submit your experiment data. This button will submit your HIT for approval and return you to Mechanical Turk.

[Submit Results]

(page break) **Almost done!**

The experiment is now completed. Thank you very much for your participation!

Your total score in the prediction section was [ ].

Base payment: [ ]

Bonus payment: [ ]

You will receive your payments within five days. Bonus payments may vary by +/- one cent due to rounding. Make sure to save this page for your records. If you have any questions, please feel free to contact us.

More Information

In case you are curious about the statistical questions at the end of the experiment, here are the answers. Your answers to these questions do not affect your payments or the quality of your performance in this HIT.

Q. What is the median of the following numbers: 10, 30, 60, 70, 90, 150, 220, 760?

A: The median is (70 + 90) / 2 = 80.

Q. A town has two hospitals. The larger hospital has on average 35 babies born every day. The smaller hospital has on average 10 babies born every day. We know that about 50 percent of babies are boys. For a period of 6 months, the hospitals recorded the number of days when more than 70 percent of the babies born are boys, and called them “baby boy days." Which of the following do you think is most likely?

A: The smaller hospital recorded more “baby boy days” than the larger hospital.

Q. A fair coin is tossed 6 times. What do you think about the likelihood of seeing Pattern A: H-T-H-T-T-H vs. Pattern B: H-H-H-T-T-T?

A: They are equally likely.

To help us with our research, please do not discuss or share these questions on public forums. Thank you very much for your cooperation!

[Submit HIT and Return to MTurk]

B.2 Variants

All experimental conditions in Experiment 1 and Experiment 2 described in Section 2 follow the sample experiment above, except they vary in the parameter ρ.

Several experimental conditions in Experiment 3 have some slight differences, which are explained below.

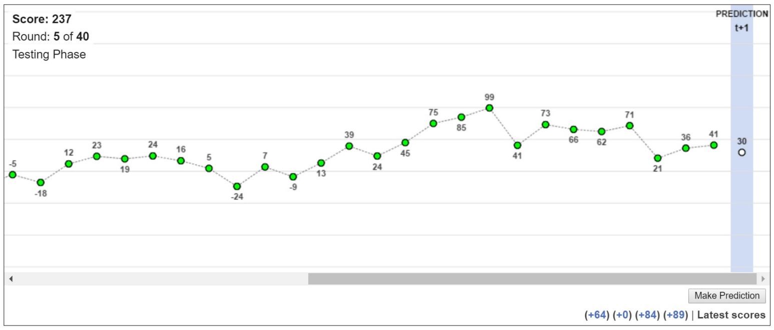
- Conditions C1 to C4 (context):
  - In the second paragraph of experimental instructions, we explain the following

    “In this study, we would like to understand how people make predictions about future realizations of random processes. **The process you will see has the same property as quarterly real GDP growth/monthly inflation/monthly stock returns/monthly house price growth in the US in the last three decades.** We will first show you 40 past realizations of a process, and you will make predictions of its future value for 40 rounds."
  - The parameters ρ, μ, σ are based on the properties of these actual processes.
  - Everything else is the same as the sample experiment above.
- Conditions C5 to C8 (no context, comparison):
  - The parameters ρ, μ, σ correspond to those in Conditions C1 to C4.

  - Everything else is the same as the sample experiment above.

- Condition C9 (comparison):
  - Everything else is the same as the sample experiment above. $\rho = 0.4$.

- Conditions C10 to C13 (forecast next realization F1 only):
  - Only forecast the next realization (instead of the next two realizations). Below is a screenshot.

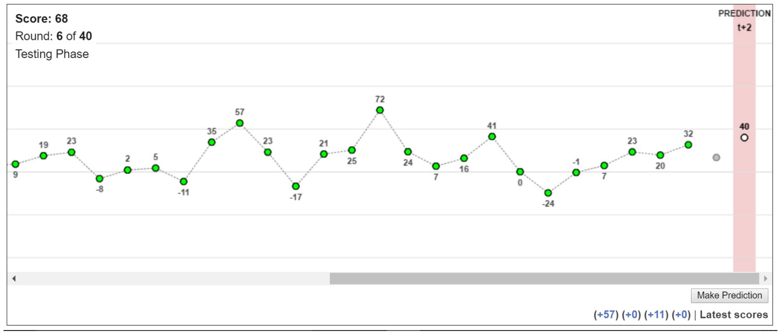


Score: 237
Round: 5 of 40
Testing Phase

(+64) (+0) (+84) (+89) | Latest scores

  - Everything else is the same as the sample experiment above.

- Conditions C14 to C17 (forecast two step ahead realization F2 only):
  - Only forecast the two step ahead realization (instead of the next two realizations). Below is a screenshot.



Score: 68
Round: 6 of 40
Testing Phase

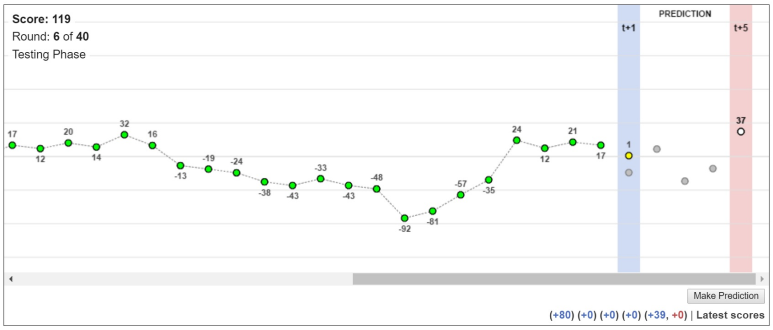
(+57) (+0) (+11) (+0) | Latest scores

  - Everything else is the same as the sample experiment above.

- Conditions C18 to C21 (forecast F1 and F5):
  - Forecast the next realization and the five step ahead realization. Below is a screenshot.
  - Everything else is the same as the sample experiment above.

- Conditions C22 to C25 (no gray dot):
  - Forecast the next two realizations, but remove the gray dot indicating $F_{t-1}x_{t+1}$. Below is a screenshot.
  - Everything else is the same as the sample experiment above.

In Experiment 4, conditions D1 and D2 have the same display as conditions in Experiment 1 and Experiment 2. In conditions D3 and D4, we include in the second paragraph of experimental instructions explains: "We will first show you 40 past realizations from 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. You will make predictions of the future value of the process for 40 rounds."

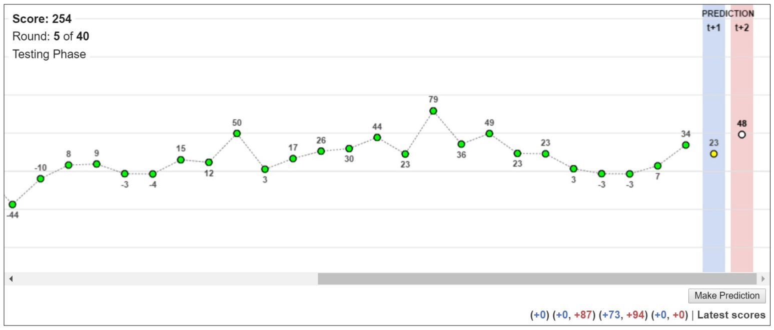


Score: 119

Round: **6** of **40**

Testing Phase

(+80) (+0) (+0) (+0) (+39, +0) | **Latest scores**



Score: 254

Round: **5** of **40**

Testing Phase

(+0) (+0, +87) (+73, +94) (+0, +0) | **Latest scores**