INTERNET APPENDIX:
OVERREACTION IN EXPECTATIONS: EVIDENCE AND THEORY*

Hassan Afrouzi, Columbia
Spencer Y. Kwon, Harvard
Augustin Landier, HEC Paris
Yueran Ma, Chicago Booth
David Thesmar, MIT Sloan

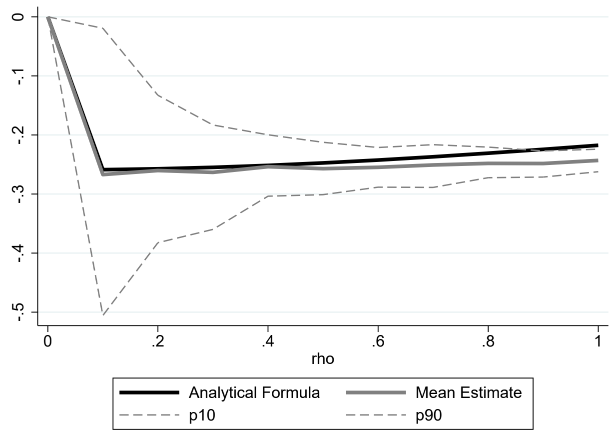
This Internet Appendix has six sections.

- Section IA1 presents additional figures and tables referenced in the main text.
- Section IA2 contains the proofs of propositions in the main text.
- Section IA3 discusses the psychological foundations of the assumptions for the model.
- Section IA4 presents an extension of the model for general ARMA processes.
- Section IA5 presents an extension of the model that incorporates underreaction into forecasts.
- Section IA6 describes additional model predictions for changing what is on top of mind.

*Emails: hassan.afrouzi@columbia.edu, ykwon@hbs.edu, landier@hec.fr, yueran.ma@chicagobooth.edu, thesmar@mit.edu.

IA1 Appendix Figures and Tables

(A) Error-Revision Coefficient



(B) Implied Persistence

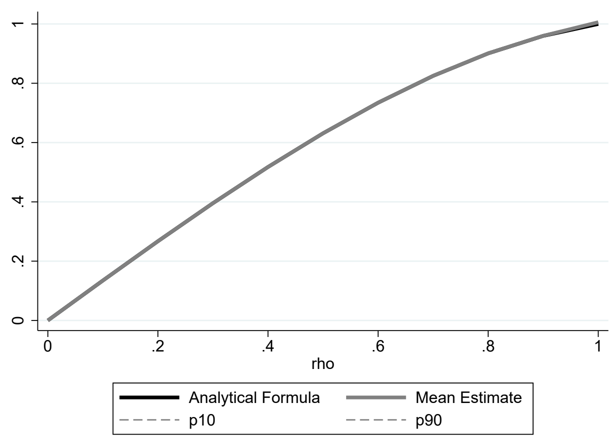


FIGURE A.1

Estimation Error: Error-Revision Coefficient and Implied Persistence

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

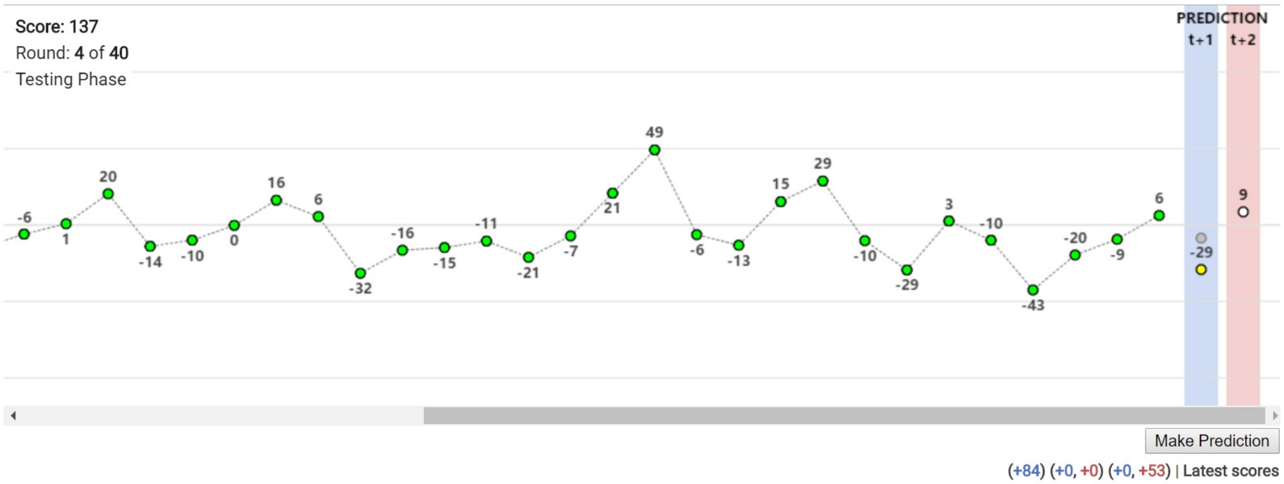


FIGURE A.2

Prediction Screen

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

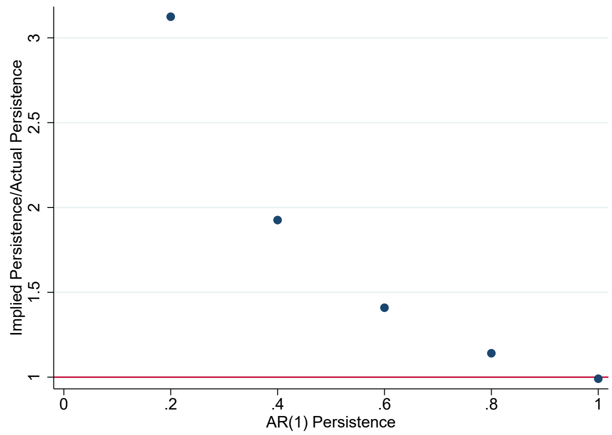
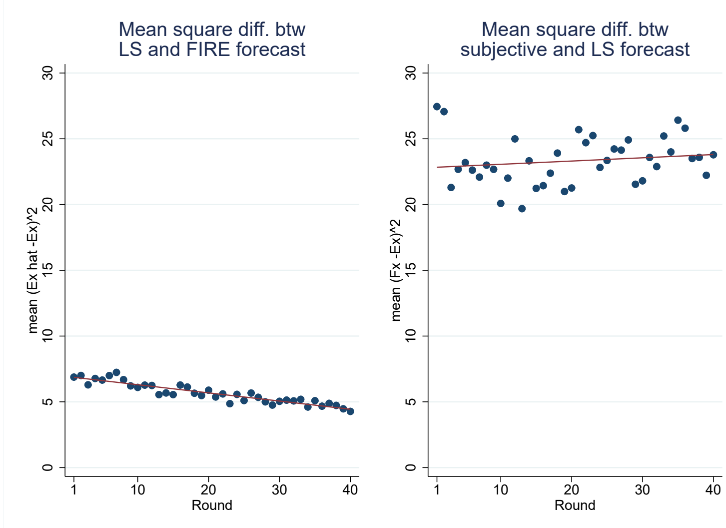


FIGURE A.3

Implied Persistence Relative to Actual Persistence

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

(A) Least Square Forecasts vs. FIRE and Subjective Forecasts



(B) Implied Persistence of Least Square Forecasts

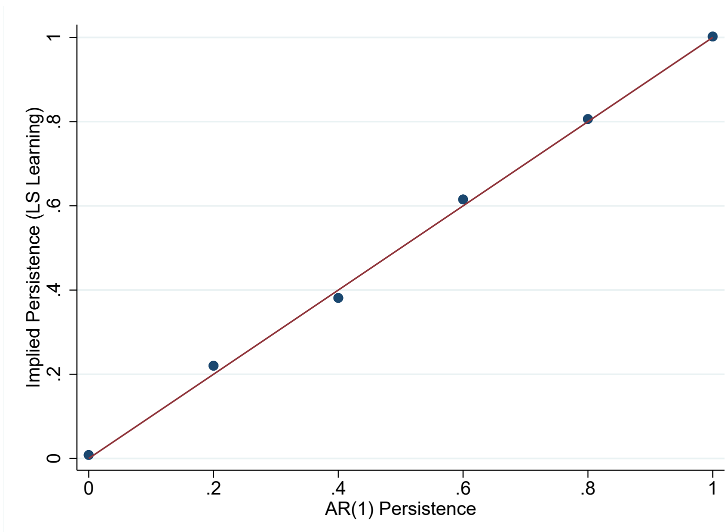
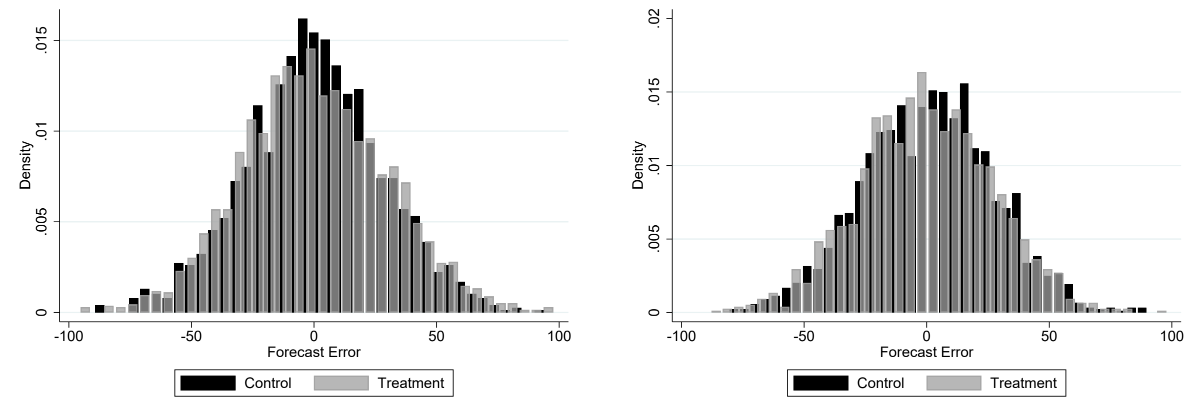


FIGURE A.4

Distance between Subjective Forecasts and Rational Expectations

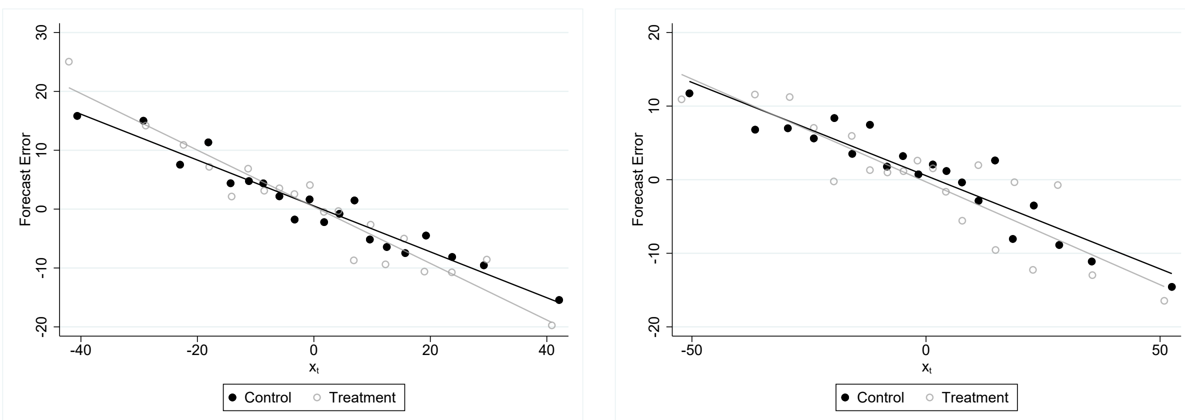
The top left panel shows the root mean squared difference between in-sample least square expectations and full information rational expectations (FIRE). The top right panel shows the root mean squared difference between participants' actual subjective forecasts and the least square forecasts. The data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

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



$\rho = 0.2$ $\rho = 0.6$

(B) Forecast Error Conditional on x_t



$\rho = 0.2$ $\rho = 0.6$

FIGURE A.5

Knowledge of Linear DGP and the Distribution of Forecasts

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

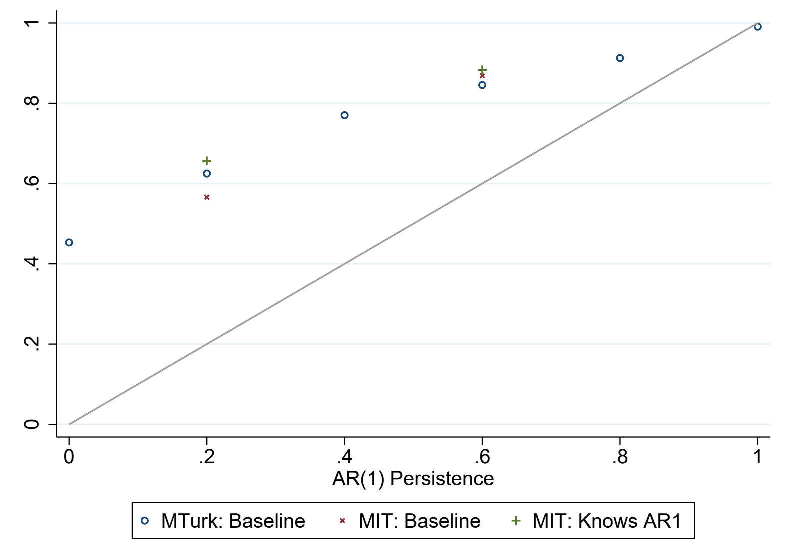
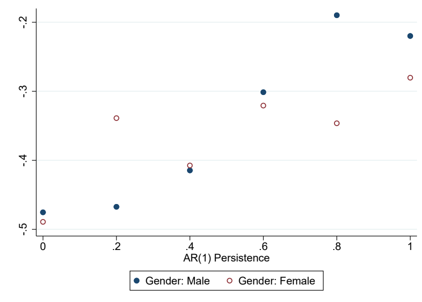


FIGURE A.6

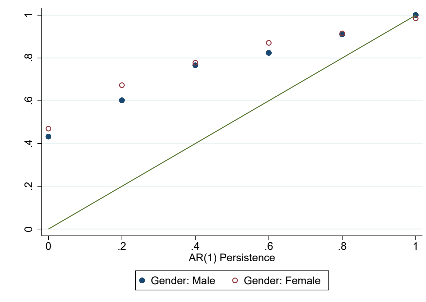
Comparison of Experiment 3 (MIT EECS) and Experiment 1 (Baseline)

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

(A) By Gender: Male vs. Female

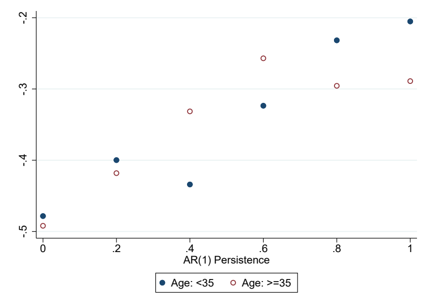


Error-Revision Coefficient by Gender

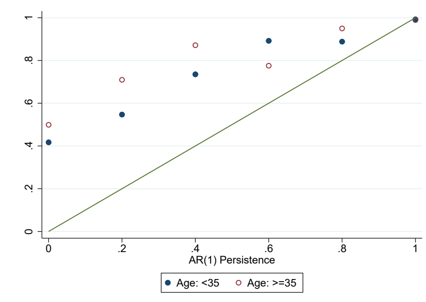


Implied Persistence by Gender

(B) By Age: Below 35 vs. Above 35

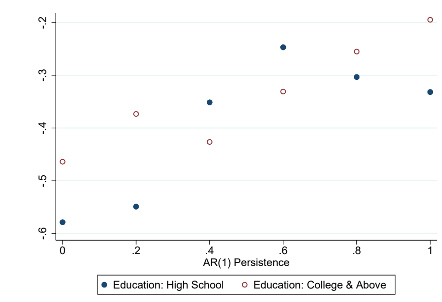


Error-Revision Coefficient by Age

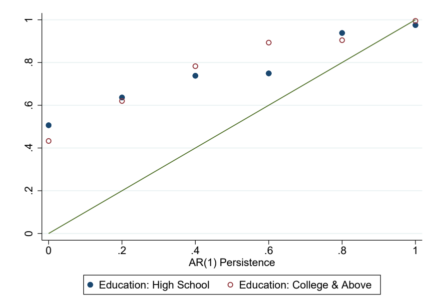


Implied Persistence by Age

(C) By Education: High School vs. College and Above



Error-Revision Coefficient by Education



Implied Persistence by Education

FIGURE A.7

Overreaction and Persistence of Process: Results by Demographics

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

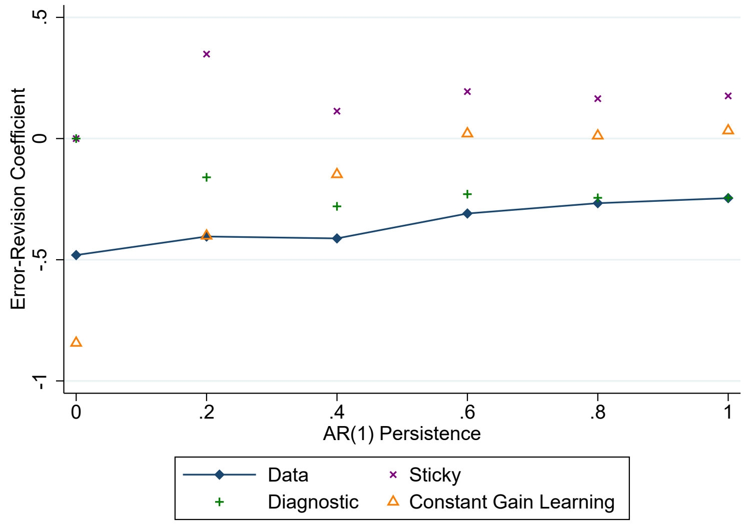


FIGURE A.8

Error-Revision Coefficient: Data vs Models

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

(A) Long-Run Mean Prediction

(B) Telling Participants the Long-Run Mean is Zero

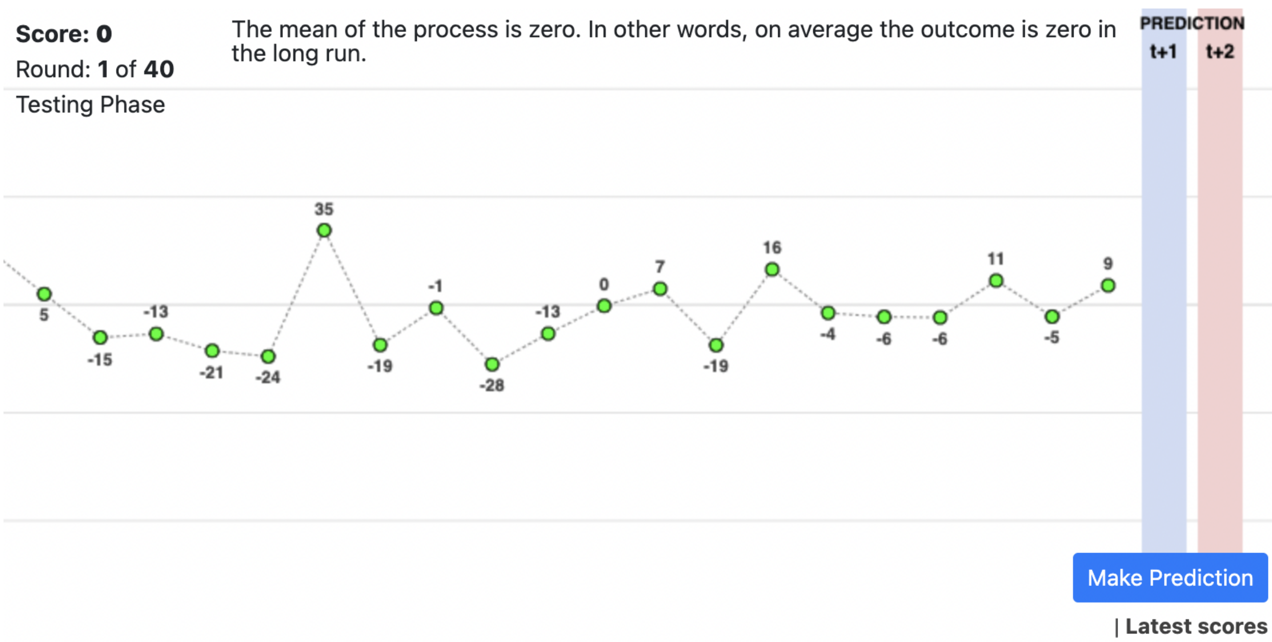


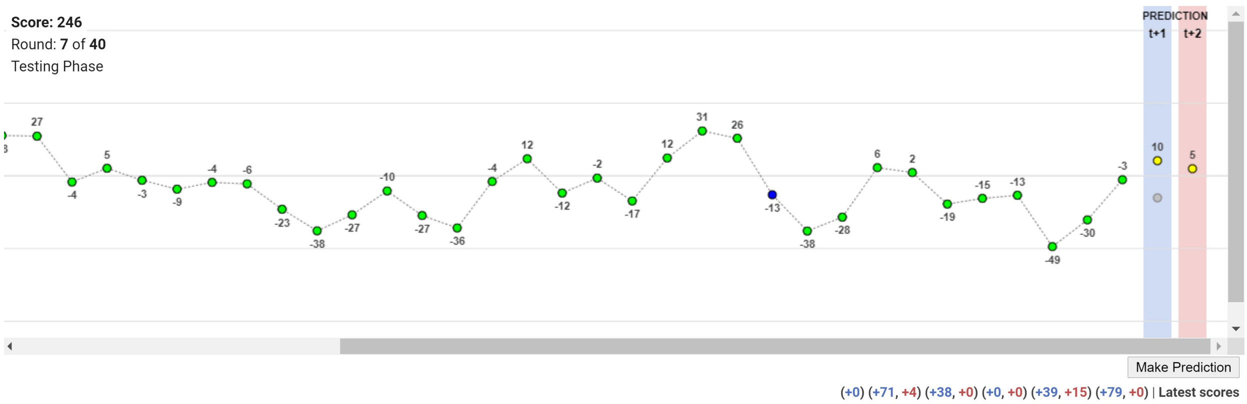
FIGURE A.9

Prediction Screen for Additional Experiments

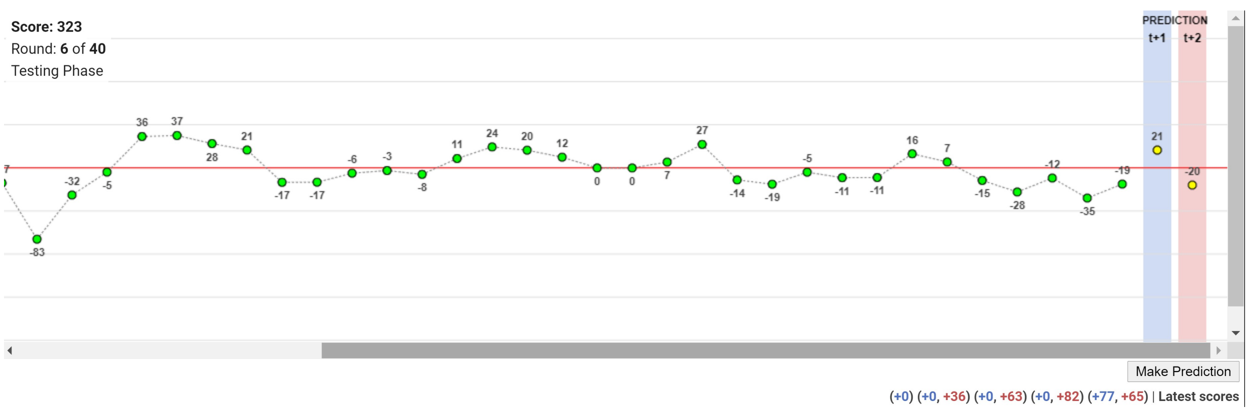
This figure shows the screenshot of the prediction task for additional conditions in Experiments 4 and 5. Panel A shows the condition where we ask participants for the long-run mean assessment. Panel B shows the condition where we tell participants that the long-run mean is zero. Panel C shows the condition where we require participants to click on x_{t-10} (the dot in blue) before making the prediction. Panel D shows the condition where we include a red line at zero.

Prediction Screen for Additional Experiments (Cont.)

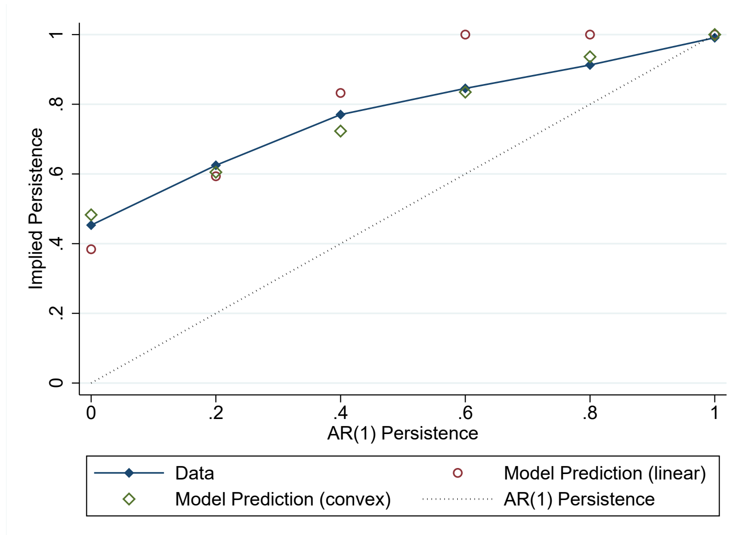
(C) Click x_{t-10}



(D) Show Red Line at 0



(A) $h = 1$



(B) $h = 10$

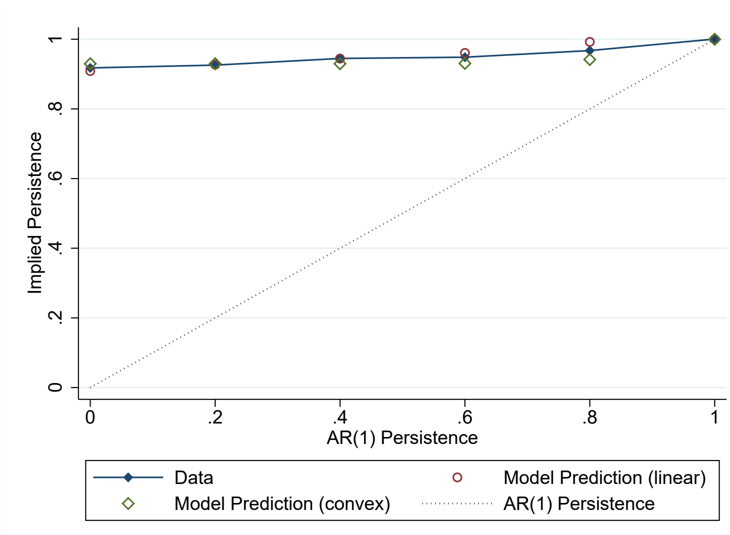
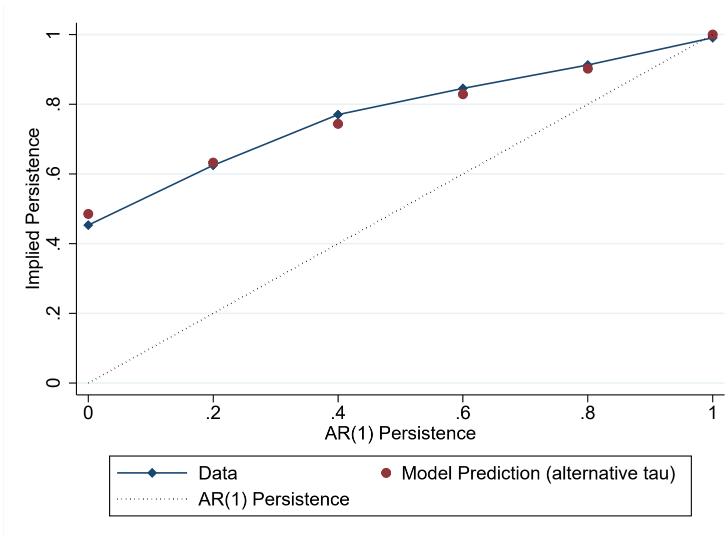


FIGURE A.10

Model Functional Form: Robustness Checks

This figure shows the model fit under alternative model specifications of the cost function, for $h = 1$ in Panel A and $h = 10$ in Panel B. The red dots represent the implied persistence from our model when $\gamma = 1$, and the green diamonds represent result from our model when we do a full grid search for γ. The blue line represents the value observed in the forecast data.

(A) $h = 1$



(B) $h = 10$

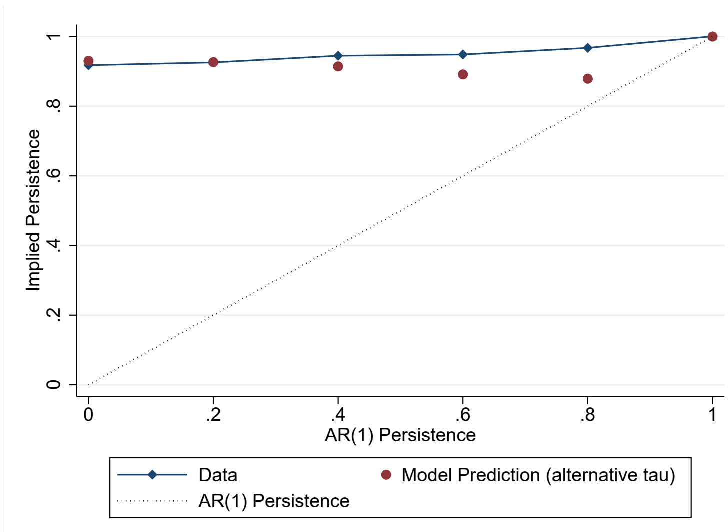


FIGURE A.11

Model Functional Form: Robustness Checks

This figure shows the model fit under the alternative formulation of $\underline{\tau}$, as discussed in Section 6.2, for $h = 1$ in Panel A and $h = 10$ in Panel B. The red dots represent the implied persistence from our model, and the blue line represents the value observed in the forecast data.

TABLE A.1

EXPERIMENTAL LITERATURE ON EXPECTATIONS FORMATION

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

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

TABLE A.2

SUMMARY OF CONDITIONS

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

Notes. This table provides a summary of the experiments we conducted. Each panel describes one experiment, and each line within a panel corresponds to one treatment condition. Columns (1) to (3) show the parameters of the AR(1) process $x_{t+1} = \mu + \rho x_t + \epsilon_{t+1}$. Experiments E41 to E46 combine participants who are randomly assigned to have high or low incentive rounds first (used in Section 6.2, so they have about twice the number of participants. Participants are only allowed to participate once.

Summary of Conditions (Continued)

	Condition	(1) Persistence ρ	(2) Mean μ	(3) Conditional Vol σ_ϵ	(4) # of Participants
Panel E: Experiment 5 – Additional Tests, MTurk					
E11	F10 + Long-Run Average	0	0	20	31
E12	F10 + Long-Run Average	0.2	0	20	32
E13	F10 + Long-Run Average	0.4	0	20	26
E14	F10 + Long-Run Average	0.6	0	20	31
E15	F10 + Long-Run Average	0.8	0	20	34
E16	F10 + Long-Run Average	1	0	20	32
E21	Baseline	0	0	20	37
E22	Baseline	0.2	0	20	40
E23	Baseline	0.4	0	20	25
E24	Baseline	0.6	0	20	30
E25	Baseline	0.8	0	20	31
E26	Baseline	1	0	20	42
E31	Inform Mean is Zero	0	0	20	36
E32	Inform Mean is Zero	0.2	0	20	31
E33	Inform Mean is Zero	0.4	0	20	33
E34	Inform Mean is Zero	0.6	0	20	28
E35	Inform Mean is Zero	0.8	0	20	42
E36	Inform Mean is Zero	1	0	20	42
E41	High vs Low Incentive	0	0	20	66
E42	High vs Low Incentive	0.2	0	20	72
E43	High vs Low Incentive	0.4	0	20	73
E44	High vs Low Incentive	0.6	0	20	69
E45	High vs Low Incentive	0.8	0	20	72
E46	High vs Low Incentive	1	0	20	67

TABLE A.3

SUMMARY STATISTICS

(A) PARTICIPANT DEMOGRAPHICS

	Experiment 1		Experiment 2		Experiment 3		Experiment 4		Experiment 5	
	Obs.	%	Obs.	%	Obs.	%	Obs.	%	Obs.	%
Gender: Female	90	43.5	61	47.7	116	56.9	316	43.1	417	40.8
Gender: Male	117	56.5	67	52.3	88	43.1	418	56.9	605	59.2
Age: <= 25	30	14.5	18	14.1	197	96.6	62	8.4	138	13.5
Age: 25-45	138	66.7	89	69.5	7	3.4	500	68.1	679	66.4
Age: 45-65	35	16.9	20	15.6	0	0.0	156	21.3	193	18.9
Age: 65+	4	1.9	1	0.8	0	0.0	16	2.2	12	1.2
Education: Grad School	20	9.7	18	14.1	0	0.0	170	23.2	353	34.5
Education: College	132	63.8	74	57.8	204	100.0	426	58.0	589	57.6
Education: High School	55	26.6	36	28.1	0	0.0	133	18.1	66	6.5
Education: Below/Other	0	0.0	0	0.0	0	0.0	5	0.7	14	1.4
Invest. Exper.: Extensive	7	3.4	3	2.3	2	1.0	77	10.5	306	29.9
Invest. Exper.: Some	58	28.0	29	22.7	21	10.3	258	35.1	441	43.2
Invest. Exper.: Limited	71	34.3	56	43.8	43	21.1	232	31.6	163	15.9
Invest. Exper.: None	71	34.3	40	31.3	138	67.6	167	22.8	112	11.0
Taken Stat Class: No	117	56.5	80	62.5	0	0.0	361	49.2	292	28.6
Taken Stat Class: Yes	90	43.5	48	37.5	204	100.0	373	50.8	730	71.4

Notes. Panel A describes demographics of participants. Panel B reports basic experimental statistics, including the total score, the total bonus (incentive payments) paid in US dollars, the overall time taken to complete the experiment, and the time taken to complete the forecasting part (the main part). The final part of Panel B separates conditions E41 to E46 where each participant makes 80 rounds of forecasts (40 with high incentives and 40 with low incentives) used in Section 6.2.

Summary Statistics (Continued)

(B) Experimental Statistics

	Mean	p25	p50	p75	SD	*N*
Experiment 1						
Total Forecast Score	2,004	1,690	1,990	2,335	461.93	207
Bonus ($)	3.34	2.82	3.32	3.89	0.77	207
Total Time (min)	18.01	10.92	13.11	21.85	11.34	207
Forecast Time (min)	6.80	4.54	5.66	7.79	3.53	207
Experiment 2						
Total Forecast Score	1,843	1,588	1,820	2,138	463.38	128
Bonus ($)	3.07	2.65	3.04	3.56	0.77	128
Total Time (min)	15.82	8.74	13.11	19.66	9.80	128
Forecast Time (min)	6.70	4.54	6.02	7.58	3.17	128
Experiment 3						
Total Forecast Score	2,071	1,755	2,046	2,326	429.59	204
Bonus ($)	8.63	7.31	8.53	9.69	1.79	204
Total Time (min)	18.45	6.55	10.92	13.11	37.67	204
Forecast Time (min)	8.78	4.03	5.09	7.46	19.72	204
Experiment 4						
Total Forecast Score	1,767	1,422	1,812	2,174	610.23	734
Bonus ($)	2.95	2.37	3.02	3.62	1.02	734
Total Time (min)	15.75	8.74	13.11	19.66	10.00	734
Forecast Time (min)	7.88	4.79	6.50	9.22	4.97	734
Experiment 5						
Total Forecast Score	1,815	1,053	1,713	2,304	1093.47	603
Bonus ($)	3.03	1.76	2.86	3.84	1.82	603
Total Time (min)	14.25	8.74	13.11	21.85	20.87	603
Forecast Time (min)	8.53	4.64	6.54	10.15	6.14	603
Experiment 5 (with two rounds)						
Total Forecast Score	2,984	1,993	3,079	3,979	1275.93	419
Bonus ($)	4.12	2.69	4.18	5.55	1.85	419
Total Time (min)	19.65	10.92	17.48	24.03	10.60	419
Forecast Time (min)	12.32	7.85	10.73	14.94	6.14	419

TABLE A.4

IMPLIED PERSISTENCE WITH AND WITHOUT GREY DOT

$\rho =$	0.2		0.4		0.6		0.8	
Grey Dot	Yes	No	Yes	No	Yes	No	Yes	No
x_t	0.62***	0.63***	0.77***	0.80***	0.85***	0.89***	0.91***	0.92***
	(0.06)	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.02)	(0.05)
Observations	1,280	1,680	1,440	1,080	1,560	1,360	1,120	1,160
R^2	0.29	0.19	0.36	0.38	0.46	0.60	0.73	0.71

Notes. This table presents regressions of $F_t x_{t+1}$ on x_t for different levels of actual persistence ρ in conditions with and without the grey dot. In columns (1), (3), (5), (7), we report the implied persistence (ρ_1^s) for the baseline conditions (A2, A3, A4, A5 in Table A.2). These baseline conditions contain a grey dot (to remind the forecaster of her earlier two-period forecast $F_{t-1}x_{t+1}$). In columns (2), (4), (6) and (8), we report the implied persistence (ρ_1^s) for conditions that are identical to the baseline but without the grey dot. These additional conditions are only available for ρ between 0.2 and 0.8; they are not described in Table A.2 for clarity of exposition, and will not be studied in the rest of the paper. Standard errors clustered by participant are presented in parentheses. *** indicates a 1% level of significance.

TABLE A.5

EFFECT OF KNOWING THE PROCESS IS AR(1)

	Baseline Condition	Knows AR(1)	Difference (p-value)
$\rho = 0.2$	0.56	0.65	0.14
$\rho = 0.6$	0.86	0.88	0.71

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

TABLE A.6

Estimation of Expectation Models

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

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

TABLE A.7

BIAS AND LAGGED REALIZATIONS

(A) LHS IS $F_t x_{t+1}$

$\rho =$	0 (1)	0.2 (2)	0.4 (3)	0.6 (4)	0.8 (5)	1 (6)
x_t	0.45***	0.64***	0.81***	0.89***	1.03***	1.16***
	(0.05)	(0.06)	(0.04)	(0.06)	(0.05)	(0.05)
x_{t-1}	0.03	-0.06	-0.10**	-0.10*	-0.15**	-0.19***
	(0.04)	(0.04)	(0.04)	(0.05)	(0.06)	(0.05)
x_{t-2}	0.00	0.04	-0.02	0.02	-0.01	-0.07*
	(0.04)	(0.03)	(0.03)	(0.05)	(0.05)	(0.04)
x_{t-3}	0.09**	0.05	0.08***	-0.01	0.03	0.06*
	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)
x_{t-4}	0.07**	0.03	-0.03	0.08**	-0.01	0.02
	(0.03)	(0.04)	(0.03)	(0.03)	(0.02)	(0.03)
Observations	1,280	1,280	1,440	1,560	1,120	1,600
R^2	0.15	0.29	0.37	0.47	0.74	0.98

(B) LHS IS $\rho x_t - F_t x_{t+1}$

$\rho =$	0 (1)	0.2 (2)	0.4 (3)	0.6 (4)	0.8 (5)	1 (6)
x_t	-0.45***	-0.44***	-0.41***	-0.29***	-0.23***	-0.16***
	(0.05)	(0.06)	(0.04)	(0.06)	(0.05)	(0.05)
x_{t-1}	-0.03	0.06	0.10**	0.10*	0.15**	0.19***
	(0.04)	(0.04)	(0.04)	(0.05)	(0.06)	(0.05)
x_{t-2}	-0.00	-0.04	0.02	-0.02	0.01	0.07*
	(0.04)	(0.03)	(0.03)	(0.05)	(0.05)	(0.04)
x_{t-3}	-0.09**	-0.05	-0.08***	0.01	-0.03	-0.06*
	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)
x_{t-4}	-0.07**	-0.03	0.03	-0.08**	0.01	-0.02
	(0.03)	(0.04)	(0.03)	(0.03)	(0.02)	(0.03)
Observations	1,280	1,280	1,440	1,560	1,120	1,600
R^2	0.15	0.16	0.13	0.08	0.06	0.03

Notes. This table shows regressions of the forecast $F_t x_{t+1}$ (Panel A) or deviation from the rational benchmark $\rho x_t - F_t x_{t+1}$ (Panel B) on lags of realizations x_{t-k}. The data comes from Experiment 1. Standard errors clustered by participant are presented in parentheses. *** indicates a 1% level of significance.

TABLE A.8

MODEL FIT

Forecast Horizon	$h=1$		$h=2$		$h=5$		$h=10$	
MSE Type	ρ_h^s	Forecast	ρ_h^s	Forecast	ρ_h^s	Forecast	ρ_h^s	Forecast
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Current model	0.002	496.0	0.001	719.5	0.001	689.9	0.000	2203.6
Adaptive	0.035	495.7	.	.	.	.	.	.
Extrapolative	0.064	527.3	.	.	.	.	.	.
Sticky	0.117	556.2	0.140	786.1	0.197	814.6	0.307	2334.1
Diagnostic	0.069	521.2	0.115	758.0	0.177	803.3	0.299	2370.2
Constant gain	0.066	526.6	0.040	750.0	0.033	735.0	0.022	2459.6

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

TABLE A.9

TELLING PARTICIPANTS THE LONG-RUN MEAN IS ZERO

	$x_{t+1} - F_t x_{t+1}$		$\rho x_t - F_t x_{t+1}$	
	(1)	(2)	(3)	(4)
x_t	-0.329***	-0.358***	-0.334***	-0.331***
	(0.041)	(0.040)	(0.037)	(0.035)
$\times$ Show Long-Run Mean $= 0$	0.106*	0.068	0.103*	0.067
	(0.060)	(0.055)	(0.057)	(0.050)
Participant FE	N	Y	N	Y
Observations	13,320	13,320	13,320	13,320
R^2	0.02	0.32	0.03	0.39

Notes. In this table, we regress different definitions of the forecast error on the last realization, interacted with indicator variables that are equal to one in treatment conditions where we tell participants that the long-run mean is equal to zero. The data is collected in Experiment 5. Each participant is randomly assigned to a given ρ and a given condition. Columns (2) and (4) include participant fixed effects to control for average optimism. In all regressions, we exclude conditions for which $\rho = 1$, since in this case we know that forecasts do not have significant biases (the implied persistence is close to one). Standard errors clustered by participant are presented in parentheses. *** indicates a 1% level of significance.

TABLE A.10

EFFECT OF HIGH INCENTIVES

(A) EFFECT ON OVERREACTION

	$x_{t+1} - F_t x_{t+1}$		$\rho x_t - F_t x_{t+1}$	
	(1)	(2)	(3)	(4)
x_t	-0.301***	-0.309***	-0.297***	-0.287***
	(0.032)	(0.028)	(0.031)	(0.026)
$x_t \times$ High Incentive	-0.023	-0.009	-0.034	-0.017
	(0.035)	(0.028)	(0.034)	(0.027)
Participant FE	N	Y	N	Y
Observations	28,160	28,160	28,160	28,160
R^2	0.03	0.27	0.03	0.33

(B) EFFECT ON FORECASTING SCORES

	Score		Log Score	
	(1)	(2)	(3)	(4)
High Incentive	10.730	10.730	0.007	-0.004
	(30.788)	(43.541)	(0.032)	(0.044)
Participant FE	N	Y	N	Y
Observations	704	704	702	702
R^2	0.00	0.82	0.00	0.80

Notes. This table analyzes the effect of high incentives. Data come from Experiment 5. Half of the participants see 40 high incentive rounds followed by 40 low incentive rounds, and half of the participants see the reverse order (the parameters for the process remain the same for the high and low incentive rounds). Participants are randomly assigned to one of these two settings and a given ρ. In Panel A, we examine how incentives affect the predictability of forecast errors. Columns (2) and (4) include participant fixed effects to control for average optimism. In Panel B, we examine how incentives affect forecasting scores. For each participant, we calculate the total score from the 40 forecasting rounds with high incentives and that from the 40 forecasting rounds with low incentives (so each participant contributes two observations). In all regressions, we exclude conditions for which $\rho = 1$, since in this case, we know that forecasts do not have significant biases (the implied persistence is close to one). Standard errors clustered by participant are presented in parentheses. *** indicates a 1% level of significance.

IA2 Proofs

IA2.1 Standard Errors of Error-Revision Coefficient

Proposition 1. Assume a univariate regression of centered variables:

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

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

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

Proof. The OLS estimator of β is given by:

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

Hence,

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

By virtue of the central limit theorem, we have:

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

while

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

This ensures that:

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

Note that the asymptotic variance can be rewritten as:

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

□

Evidently, this ratio is bigger when the variance of x_i is smaller.

For the error-revision coefficient, it can easily be shown that:

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

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

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

IA2.2 Lemma 1

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

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

$$F_t^* x_{t+h}(S_t) \equiv \arg\min_{F_t x_{t+h}} \mathbb{E}[(F_t x_{t+h} - x_{t+h})^2 | S_t] \Rightarrow F_t^* x_{t+h}(S_t) = \mathbb{E}[x_{t+h}|S_t]. \tag{IA2.1}$$

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

$$\mathbb{E}[(F_t^* x_{t+h}(S_t) - x_{t+h})^2 | S_t] = \mathrm{var}(x_{t+h}|S_t). \tag{IA2.2}$$

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

$$\mathrm{var}(x_{t+h}|S_t) = \mathrm{var}((1-\rho^h)\mu + \rho^h x_t + \sum_{j=1}^{h} \rho^{h-j} \varepsilon_{t+j} | S_t) \tag{IA2.3}$$

$$= (1-\rho^h)^2 \mathrm{var}(\mu|S_t) + \sigma_\varepsilon^2 \sum_{j=1}^{h} \rho^{2(h-j)}, \tag{IA2.4}$$

where the second line follows from:

1. orthogonality of future innovations to S_t that follows from feasibility ($\varepsilon_{t+j} \perp \mathcal{A}_t, \forall j \geq 1$);

2. $\mathrm{var}(x_t|S_t) = 0$ since $x_t \in S_t$ by assumption.

It is important to note that the second term in Equation IA2.4 is independent of the choice for S_t. We can now rewrite the agent's problem as:

$$\min_{S_t} \mathbb{E}[(1-\rho^h)^2 \text{var}(\mu|S_t) + C(S_t)|x_t] \tag{IA2.5}$$

$$s.t. \{x_t\} \subseteq S_t \subseteq \mathcal{A}_t, \tag{IA2.6}$$

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

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

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

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

$$C(\hat{S}_t) \leq C(S_t) \Leftrightarrow \mathbb{I}(\hat{S}_t, x_{t+h}|x_t) \leq \mathbb{I}(S_t, x_{t+h}|x_t). \tag{IA2.7}$$

A final observation yields our desired result: by definition of the mutual information function in terms of entropy,[1]

$$\mathbb{I}(S_t; \mu|x_t) = h(\mu|x_t) - \mathbb{E}[h(\mu|S_t)|x_t]. \tag{IA2.8}$$

Similarly,

$$\mathbb{I}(\hat{S}_t; \mu|x_t) = h(\mu|x_t) - \mathbb{E}[h(\mu|\hat{S}_t)|x_t]. \tag{IA2.9}$$

It follows from these two observations that

$$C(\hat{S}_t) \leq C(S_t) \Leftrightarrow \mathbb{E}[h(\mu|\hat{S}_t)|x_t] \geq \mathbb{E}[h(\mu|S_t)|x_t]. \tag{IA2.10}$$

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

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

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

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

$$\mathbb{I}(\mu; S_t|x_t) = h(\mu|x_t) - h(\mu|S_t) \tag{IA2.11}$$
$$= \frac{1}{2\ln(2)}\ln(\text{var}(\mu|x_t)) - \frac{1}{2\ln(2)}\ln(\text{var}(x_t|S_t)). \tag{IA2.12}$$

Moreover, for simplicity, we define $\tau(S_t) \equiv \text{var}(\mu|S_t)^{-1}$ as the precision of belief about μ generated by S_t and $\underline{\tau} \equiv \text{var}(\mu|x_t)^{-1}$ as the precision of the prior belief of the agent about μ. Moreover, for ease of notation and without loss of generality, we normalize the mutual information function by the constant $2\ln(2)$.[3] It follows that

$$\mathbb{I}(\mu; S_t|x_t) = \ln\left(\frac{\tau(S_t)}{\underline{\tau}}\right), \tag{IA2.13}$$
$$C(S_t) = \omega\frac{\exp(\gamma \cdot \mathbb{I}(\mu; S_t|x_t)) - 1}{\gamma} \tag{IA2.14}$$
$$= \omega\frac{\left(\frac{\tau(S_t)}{\underline{\tau}}\right)^{\gamma} - 1}{\gamma}. \tag{IA2.15}$$

Hence, the agent's problem can be rewritten as

$$\min_{S_t} \mathbb{E}\left[\frac{(1-\rho^h)^2}{\tau(S_t)} + \omega\frac{\left(\frac{\tau(S_t)}{\underline{\tau}}\right)^{\gamma} - 1}{\gamma}\Bigg| x_t\right] \tag{IA2.16}$$
$$s.t. \{x_t\} \subseteq S_t \subseteq \mathcal{A}_t. \tag{IA2.17}$$

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

The proof follows from optimizing over the PDF of the distribution of X:

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

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

IA2.3 Proposition 1

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

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

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

$$\text{var}(\mu|S_t) = \tau^{-1} = \underline{\tau}^{-1} \min\left\{1, \left(\frac{\omega\underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}\right\}. \tag{IA2.18}$$

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

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

$$\mu_t \equiv \mathbb{E}[\mathbb{E}[x_{t+h}|S_t]|\mu, x_t] = (1-\rho^h)\mathbb{E}[\mathbb{E}[\mu|S_t]|\mu, x_t] + \rho^h\mathbb{E}[\mathbb{E}[x_t|S_t]|\mu, x_t] = x_t \tag{IA2.19}$$

2. if $\left(\frac{\omega\underline{\tau}}{(1-\rho^h)^2}\right) < 1$, then it means that the agent utilizes more information than what is revealed by the context x_t. Suppose a signal structure $\tilde{S}_t$ generates this posterior variance.[4] By Lemma 1 this has to be Gaussian. First, it is convenient to observe that the set $\hat{S}_t \equiv \{x_t, \mathbb{E}[\mu|\tilde{S}_t]\}$ is a sufficient statistic for $\tilde{S}_t$. To see the equivalence of the two sets, note that both are comprised of Gaussian variables and by the law of total variance, both sets generate the same posterior variance for the agent.[5]

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

[5] Formally, we have

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

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

Now, by Bayesian updating for Gaussians:

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

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

$$\text{cov}(\mu, \mathbb{E}[\mu|\tilde{S}_t]|x_t) = \text{var}(\mathbb{E}[\mu|\tilde{S}_t]|x_t) = \underline{\tau}^{-1} - \tau^{-1}, \tag{IA2.20}$$

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

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

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

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

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

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

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

$$\begin{aligned} &\mathbb{E}[\mathbb{E}[\mu|\tilde{S}_t]|\mu, x_t] = (1 - \frac{\underline{\tau}}{\tau})\mu + \frac{\underline{\tau}}{\tau}x_t \\ \Rightarrow &\mu_t \equiv \mathbb{E}[\mathbb{E}[x_{t+h}|\tilde{S}_t]|\mu, x_t] = (1-\rho^h)(1 - \frac{\underline{\tau}}{\tau})\mu + (1-\rho^h)\frac{\underline{\tau}}{\tau}x_t + \rho^h x_t. \end{aligned} \tag{IA2.21}$$

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

$$\mu_t = E_t x_{t+h} + (1-\rho^h)\frac{\underline{\tau}}{\tau}(x_t - \mu) \tag{IA2.22}$$

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

Finally, define $u_t \equiv F_t x_{t+h} - \mu_t$. Plugging in the expression for τ from (IA2.18) into (IA2.22), and setting (normalizing) $\mu = 0$, we get the expression of interest:

$$F_t x_{t+h} = \mu_t + u_t = E_t x_{t+h} + (1-\rho^h)\min\left\{1, \left(\frac{\omega\underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}\right\} x_t + u_t \tag{IA2.23}$$

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

IA2.4 Proposition 2

Proof. From Proposition 1 we can derive Δ as

$$\Delta = (1-\rho^h)\min\left\{1, \left(\frac{\omega\underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}\right\}. \tag{IA2.24}$$

1. Note that if $\Delta = 0$ then either $\rho^h = 1$ or $\omega = 0$, but recall that this expression for the precision of the long-run mean was derived under the assumption that $\text{var}(\mu|x^t)$ is arbitrarily small. So $\Delta = 0$ if and only if either $\rho = 1$ or $\omega = 0$ and past information potentially available to the forecaster is infinite.

2. As long as $\gamma \geq 0$, which is true by assumption, it is straightforward to verify that Δ is increasing in ω and $\underline{\tau}$.

3. For Δ to be decreasing in ρ^h it has to be the case that $(1-\rho^h)^{1-\frac{2}{1+\gamma}}$ is decreasing in ρ^h, which is the case if and only if

$$1 - \frac{2}{1+\gamma} \geq 0 \Leftrightarrow \gamma \geq 1. \tag{IA2.25}$$

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

$$\ln(\zeta(\rho,h)) = \ln\left(1 + (\rho^{-h}-1)\min\left\{1, \left(\frac{\omega\underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}\right\}\right). \tag{IA2.26}$$

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

□

IA2.5 Corollary 1

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

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

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

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

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

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

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

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

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

IA3 Psychological Foundations

In this section, we provide a discussion of the psychological literature on working memory (see, e.g., Baddeley, 1992) that motivates our modeling assumptions in Section 5. Our model in Section 5 has two assumptions: (1) only a subset of information is on top of mind, and such information is easier to utilize (even if all the data is in front of a participant), and (2) the most recent observation is immediately on top of mind, and other information can be actively processed by the participant with additional effort. Below, we discuss the psychological studies that relate to each of these assumptions.

1. Only a subset of information is on top of mind. A series of research in psychology on working memory emphasizes that some information is more actively utilized than others.

This notion has been referred to as heightened activation, increased accessibility, conscious awareness, or focus of attention (Baddeley and Hitch, 1993; Cowan, 1998, 2017a; Unsworth and Spillers, 2010), and is connected more broadly to costly information processing in a variety of settings (Spillers, Brewer and Unsworth, 2012). In this paper, we use the term "on top of mind" to refer to the set of information actively utilized, which corresponds to the set S_t in the model.

Our model draws on the working memory literature because overreaction to some observations suggests that forecasters use such information more than others. Indeed, a recent survey paper by Cowan (2017a) explains the concept of working memory as "the ensemble of components of the mind that hold information *temporarily* in a *heightened state of availability* for use in ongoing *information processing*."[6]

While the working memory literature provides a context for why some information has heightened utilization, a key question for our application is whether this concept applies to our experiment. Since a number of observations are shown on the experimental screen, one might wonder if working memory is a binding constraint on participants' information processing. The working memory literature highlights that heightened activation is not necessarily about recalling previously seen observations but rather about focusing on a subset of the available information in processing. Unsworth and Spillers (2010) provide a review of relevant studies. Several of them study visual tasks, which are closest to our forecasting experiment (Bleckley, Durso, Crutchfield, Engle and Khanna, 2003; Poole and Kane, 2009). To summarize the evidence, Unsworth and Spillers (2010) write: "Baddeley (1993) noted that 'the central executive component of working memory does not itself involve storage, which produces the apparently paradoxical conclusion that not all working memory studies need involve memory' (p. 167)."

Finally, a remaining question is how small is the set of constrained focus (S_t) relative to all the information that is available on the screen? Numerous experimental studies in the psychology literature also finds that the capacity of working memory is small, with a limited amount of information in a rapidly accessible state (Cowan, 2010, 2012). For a more detailed treatment of this concept, see, e.g., the book titled "working memory capacity" (Cowan, 2012); in Chapter 1, Cowan defines working memory capacity as "[the] relatively small amount of information that one can hold in mind, attend to, or, technically speaking, maintain in a rapidly accessible state, at one time," a definition that we have used as a guideline for modeling the information set that is available for processing. Later, in Chapter 4, Cowan provides evidence and makes a case for the fact that working memory capacity is limited to at most a few items or "chunks" at any given time. Our own findings are consistent with the notion that despite the abundant availability of information on the screen, participants do not necessarily process everything.

2. The most recent observation is immediately on top of mind. The next question is what determines the small amount of information that is in a state of heightened activation (on top of mind). Here we also use the psychological literature as a guideline. In the model, we assume that the most recent observation is easiest to utilize (i.e., recency effect), while other information is incorporated into forecasts through a more deliberate process. This process

[6] Cowan (2017b) conveys the idea through its title, "Working Memory: The Information You Are Now Thinking of."

is captured by the assumption that $x_t \in S_t, \forall t \geq 0$, but that S_t can be expanded by the agent through a more conscious utilization process that comes at a cognitive cost. In other words, agents in our model display immediate and automatic recency, modulated by the effortful processing of further information.

Our modeling approach is based on the psychology literature's view of the two ways for information to be incorporated in working memory, perhaps best summarized by Hitch, Hu, Allen and Baddeley (2018): "Previous research ... indicates that access to the focus of attention (FoA) can be achieved in either of two ways. The first is automatic and is indexed by *the recency effect*, the enhanced retention of the final item. The second is strategic and based on instructions to prioritize items differentially, a process that draws on executive capacity and boosts retention of information deemed important." These two forces in the working memory mechanism correspond to broader themes in psychology research about information processing: recency (which is part of the information that is automatically activated) and goal-driven information processing (which is slower and requires effort).

First, the psychology literature has studied the recency effect and how recent observations have heightened activation since the 1960s (Sternberg, 1966). Baddeley (2007); Baddeley and Hitch (1993) provide a comprehensive summary of findings about recency effects in psychology and its relationship with the mechanism of working memory. As Baddeley and Hitch (1993) write: the "presentation and processing of an item results in the activation of its node... [and] the recency effect occurs because recently activated nodes are easy to reactivate," and conclude that "the recency effect can be viewed as reflecting the utilization of automatic activation by an active, multi-component, working-memory system."[7] Experiments of working memory and recency in a visual setting are closest to our experimental framework (see, e.g., Hay, Smyth, Hitch and Horton, 2007; Phillips and Christie, 1977). In sum, the literature on recency shows that recent information gets processed automatically and enjoys heightened activation.

Second, the working memory mechanism also features the role of executive capacity and the slower goal-oriented processing of additional information. As Hu, Allen, Baddeley and Hitch (2016) summarize, "what is accessible in working memory reflects both top-down, goal-driven priorities under executive control and the results of automatic perceptual selection from the external environment." In particular, by varying which information is goal-relevant, a series of working-memory experiments demonstrate how these two forces shape the limited amount of information with heightened activation (Hitch et al., 2018).

The two components of the working memory mechanism summarized above can also be viewed as a microfoundation for a broad class of dual process models in psychology (Barrett, Tugade and Engle, 2004). The dual process models are unified by a framework where the individual starts from a default driven by what is immediately accessible ("System 1"), and further adjusts beliefs by effortful processing ("System 2"). See Evans (2008) for a summary of the many types of dual process models in the psychology literature and Ilut and Valchev (2023) for an application of dual process in economics.

Taken together, our model incorporates these two types of forces: agents automatically form a prior based on the recent observation (automatic recency), and consciously processed

[7] The working memory literature's view on recency effect has evolved over the last few decades. Earlier research on working memory often treated recency as a separate topic (Baddeley and Hitch, 1993). However, later work views that recency influences what information gets used automatically without conscious processing.

additional data (deliberate processing). More generally, the framework of our model can accommodate various settings where the forecasters are influenced by recency but may also exert costly effort to utilize more past information. Such recency effects can arise due to psychological forces as described above; they could also come from institutional forces that limit the usage of past data.

IA4 Generalized Model for ARMA Processes

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

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

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

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

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

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

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

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

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

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

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

where ($\succeq 0$) denotes positive-semidefiniteness. This is a convex optimization problem on the *positive semi-definite cone*, similar to the problem studied in Afrouzi and Yang (2021). While Afrouzi and Yang (2021) only considers the case of $\gamma \to 0$, we solve for the more general case of $\gamma > 0$. Since the cost of inaccuracy approaches infinity if $|\Sigma| \to 0$, the optimal subjective variance Σ should have a strictly positive determinant, with all the eigenvalues of

Σ strictly positive ($\Sigma \succ \mathbf{0}$). In other words, we can ignore the constraint $\Sigma \succ 0$ as it should not bind under the solution. On the other hand, the constraint $\Sigma \preceq \underline{\Sigma}$, however, potentially binds and needs to be considered (this intuitively corresponds to the case in which zero costly learning occurs).

We assume Λ is the generalized Lagrange multiplier on this constraint. It follows from convex optimization that Λ is also positive semi-definite, commutes with $X \equiv \underline{\Sigma} - \Sigma$, and satisfies complementarity slackness $\Lambda X = X\Lambda = \mathbf{0}$ (See Afrouzi and Yang (2021) for details). The first order condition is then

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

which can be rewritten as

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

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

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

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

$$\hat{\Omega}\hat{X} = \hat{\Omega} - \hat{\omega}I + \hat{\Lambda}. \tag{IA4.1}$$

Note that $\hat{X}\hat{\Lambda} = \hat{\Lambda}\hat{X} = 0$. We can also see that $\hat{\Omega}\hat{X} = \hat{X}\hat{\Omega}$ since the right hand side of Equation (IA4.1) above is symmetric. Finally, we can see that $\hat{\Lambda}$ and $\hat{\Omega}$ also commute.[8] Thus, since $\hat{\Omega}$, $\hat{X}_t$ and $\hat{\Lambda}$ are all symmetric, they are all diagonalizable, and given that they all commute with one another, they must be simultaneously diagonalizable. This implies that there are diagonal matrices D_{Λ}, D_X and D_{Ω}, as well as an orthonormal basis U ($UU' = U'U = I$), such that

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

Now multiplying Equation (IA4.1) by U from left and U' from right, we have

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

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

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

or in matrix form:

$$I - \hat{X} = \max\{\frac{\hat{\Omega}}{\hat{\omega}}, I\}^{-1} = \max\{\frac{\underline{\Sigma}^{\frac{1}{2}}\Omega\underline{\Sigma}^{\frac{1}{2}}}{\hat{\omega}}, I\}^{-1}, \tag{IA4.2}$$

[8] To see this, multiply the Equation (IA4.1) by $\hat{\Lambda}$ form right and note that $\hat{\Omega}\hat{\Lambda}$ has to be symmetric, indicating that $\hat{\Lambda}\hat{\Omega} = (\hat{\Lambda}\hat{\Omega})' = \hat{\Omega}\hat{\Lambda}$.

or

$$\Sigma = \underline{\Sigma}^{\frac{1}{2}} \max\{\frac{\underline{\Sigma}^{\frac{1}{2}}\Omega\underline{\Sigma}^{\frac{1}{2}}}{\hat{\omega}}, I\}^{-1}\underline{\Sigma}^{\frac{1}{2}}, \tag{IA4.3}$$

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

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

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

Thus, taking the log-determinant of Equation (IA4.2) (which is permitted because both sides are strictly positive definite) gives:

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

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

$$\log(\hat{\omega}) = \log(\omega) + \gamma \sum_{\lambda_i \geq \hat{\omega}} \log(\frac{\lambda_i}{\hat{\omega}}). \tag{IA4.4}$$

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

To prove the existence of a solution, note that the left hand side is increasing in $\hat{\omega}$ and subjects onto all of $\mathbb{R}$. On the other hand, the right hand side is decreasing in $\hat{\omega}$, with its range being $[\log(\omega), \infty)$. Thus, there is a unique $\hat{\omega}$ that solves this equation (which incidentally is larger than ω for $\gamma > 0$ as long as there is at least one eigenvalue larger than ω). Thus Equations (IA4.3) and (IA4.4) together pin down the optimal Σ for the agent. Therefore, applying standard Kalman filtering results, we obtain that the agent's belief about the long run mean is given by

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

where

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

and Σ is the solution in Equation (IA4.3).

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

IA5 Underreaction

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

$$s_t = x_t + \epsilon_t, \epsilon_t \sim N\left(0, \tau_\epsilon^{-1}\right). \tag{IA5.1}$$

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

$$\hat{x}_t | s_t = \frac{\tau_\epsilon}{\tau_0 + \tau_\epsilon} s_t + \frac{\tau_0}{\tau_0 + \tau_\epsilon} \bar{x}. \tag{IA5.2}$$

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

$$\hat{\mu} \sim N\left(\hat{x}_t, \underline{\tau}\right). \tag{IA5.3}$$

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

The derivations are similar as before and we have:

$$E[\mu | \hat{x}_t, S_t] = \min\left\{1, \left(\frac{\omega \underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}\right\} \hat{x}_t \tag{IA5.4}$$

$$F_t x_{t+h} = \rho^h \cdot \hat{x}_t + (1-\rho^h) \min\left\{1, \left(\frac{\omega \underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}\right\} \hat{x}_t + \underbrace{\varepsilon_t}_{\text{noise}} \tag{IA5.5}$$

$$= \rho^h x_t + \left[\underbrace{\frac{\tau_\epsilon}{\tau_0 + \tau_\epsilon}(1-\rho^h) \min\left\{1, \left(\frac{\omega \underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}\right\}}_{\text{overreaction}} - \underbrace{\frac{\tau_0}{\tau_0 + \tau_\epsilon}\rho^h}_{\text{underreaction}}\right] x_t + \text{constant} + \varepsilon_t. \tag{IA5.6}$$

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

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

Proposition 2. Holding fixed the noisy information parameters $\tau_\epsilon, \tau_0 < \infty$, there is overreaction ($\rho_h^s > \rho$) for sufficiently low ρ, and underreaction ($\rho_h^s < \rho$) if $\rho \mapsto 1$. If $\gamma \geq 1$, $\Delta = \rho_{s,h} - \rho^h$ is decreasing in ρ^h.

Proof. We have:

$$\rho_{s,h} - \rho^h = \frac{\tau_\epsilon}{\tau_0 + \tau_\epsilon}(1-\rho^h)\min\left\{1, \left(\frac{\omega\underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}\right\} - \frac{\tau_0}{\tau_0+\tau_\epsilon}\rho^h. \tag{IA5.7}$$

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

$$\frac{\tau_\epsilon}{\tau_0+\tau_\epsilon} - \rho^h, \tag{IA5.8}$$

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

$$\frac{\tau_\epsilon}{\tau_0+\tau_\epsilon}(1-\rho^h)\left(\frac{\omega\underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}} - \frac{\tau_0}{\tau_0+\tau_\epsilon}\rho^h = \frac{\tau_\epsilon}{\tau_0+\tau_\epsilon}(\omega\underline{\tau})^{\frac{1}{1+\gamma}}(1-\rho^h)^{-\frac{\gamma-1}{1+\gamma}} - \frac{\tau_0}{\tau_0+\tau_\epsilon}\rho^h. \tag{IA5.9}$$

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

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

IA6 Model Predictions for Changing What's on Top of Mind

In this section, we describe our model's predictions for the additional experiments in Section 6.1 (where we change what's on top of mind).

IA6.1 Setup

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

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

$$\mu | x_t, I \sim N(z_t, \underline{\tau} + \bar{\tau}') \tag{IA6.1}$$

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

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

$$E[\mu | x_t, S_t, I] = \min\left\{1, \left(\frac{\omega(\underline{\tau} + \bar{\tau}')}{(1-\rho)^2}\right)^{\frac{1}{1+\gamma}}\right\} z_t, \tag{IA6.2}$$

and consequently:

$$\rho^s_{1,I} = \rho + (1-\rho) \cdot \min\left\{1, \left(\frac{\omega(\underline{\tau} + \bar{\tau}')}{(1-\rho)^2}\right)^{\frac{1}{1+\gamma}}\right\} \cdot \frac{\underline{\tau}}{\underline{\tau} + \bar{\tau}'}. \tag{IA6.3}$$

In comparison, our original expression is:

$$\rho^s_1 = \rho + (1-\rho) \cdot \min\left\{1, \left(\frac{\omega\underline{\tau}}{(1-\rho)^2}\right)^{\frac{1}{1+\gamma}}\right\}. \tag{IA6.4}$$

IA6.2 Result

We have the following proposition:

Proposition 3. The implied persistence curve in the new conditions $\rho^s_{1,I}$ lies below the original implied persistence curve ρ^s_1. In other words, $\rho^s_{1,I} < \rho^s_1$ for each level of actual ρ (except $\rho = 1$).

Proof. It suffices to show:

$$\min\left\{1, \left(\frac{\omega\underline{\tau}}{(1-\rho)^2}\right)^{\frac{1}{1+\gamma}}\right\} > \min\left\{1, \left(\frac{\omega(\underline{\tau} + \bar{\tau}')}{(1-\rho)^2}\right)^{\frac{1}{1+\gamma}}\right\} \cdot \frac{\underline{\tau}}{\underline{\tau} + \bar{\tau}'}. \tag{IA6.5}$$

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

$$\begin{aligned} \left(\frac{\omega\underline{\tau}}{(1-\rho)^2}\right)^{\frac{1}{1+\gamma}} &> \left(\frac{\omega(\underline{\tau} + \bar{\tau}')}{(1-\rho)^2}\right)^{\frac{1}{1+\gamma}} \cdot \frac{\underline{\tau}}{\underline{\tau} + \bar{\tau}'} \\ \iff \left(\frac{\underline{\tau}}{\underline{\tau} + \bar{\tau}'}\right)^{\frac{1}{1+\gamma}} &> \frac{\underline{\tau}}{\underline{\tau} + \bar{\tau}'}, \end{aligned} \tag{IA6.6}$$

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

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

$$\left(\frac{\omega\underline{\tau}}{(1-\rho)^2}\right)^{\frac{1}{1+\gamma}} > \frac{\underline{\tau}}{\underline{\tau}+\bar{\tau}'}. \tag{IA6.7}$$

This is clearly true, as:

$$\left(\frac{\omega\underline{\tau}}{(1-\rho)^2}\right)^{\frac{1}{1+\gamma}} = \left(\frac{\omega(\underline{\tau}+\bar{\tau}')}{(1-\rho)^2}\right)^{\frac{1}{1+\gamma}} \cdot \left(\frac{\underline{\tau}}{\underline{\tau}+\bar{\tau}'}\right)^{\frac{1}{1+\gamma}} > \left(\frac{\underline{\tau}}{\underline{\tau}+\bar{\tau}'}\right)^{\frac{1}{1+\gamma}} > \frac{\underline{\tau}}{\underline{\tau}+\bar{\tau}'}. \tag{IA6.8}$$

□

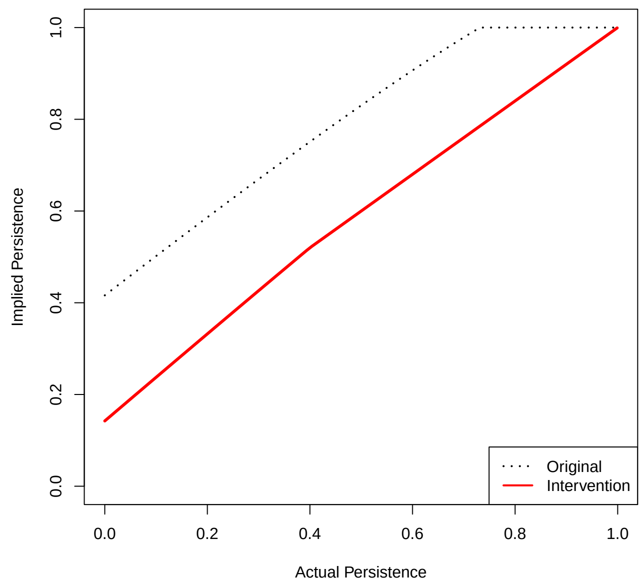


FIGURE A.12

Model Prediction for Persistence in Additional Treatment Conditions

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

References

Afrouzi, Hassan and Choongryul Yang, "Dynamic rational inattention and the Phillips Curve," CESifo Working Paper No. 8840 2021.

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

Asparouhova, Elena, Michael Hertzel, and Michael Lemmon, "Inference from streaks in random outcomes: Experimental evidence on beliefs in regime shifting and the law of small numbers," *Management Science*, 2009, *55* (11), 1766–1782.

Baddeley, Alan, "Working memory," *Science*, 1992, *255* (5044), 556–559.

— , *Working memory, thought, and action*, Vol. 45, Oxford University Press, 2007.

Baddeley, Alan D and Graham Hitch, "The recency effect: Implicit learning with explicit retrieval?," *Memory & Cognition*, 1993, *21* (2), 146–155.

Barrett, Lisa Feldman, Michele M Tugade, and Randall W Engle, "Individual differences in working memory capacity and dual-process theories of the mind," *Psychological Bulletin*, 2004, *130* (4), 553.

Beshears, John, James J Choi, Andreas Fuster, David Laibson, and Brigitte C Madrian, "What goes up must come down? Experimental evidence on intuitive forecasting," *American Economic Review*, 2013, *103* (3), 570–574.

Bleckley, M Kathryn, Francis T Durso, Jerry M Crutchfield, Randall W Engle, and Maya M Khanna, "Individual differences in working memory capacity predict visual attention allocation," *Psychonomic Bulletin & Review*, 2003, *10* (4), 884–889.

Bloomfield, Robert and Jeffrey Hales, "Predicting the next step of a random walk: Experimental evidence of regime-shifting beliefs," *Journal of Financial Economics*, 2002, *65* (3), 397–414.

Bordalo, Pedro, Nicola Gennaioli, Yueran Ma, and Andrei Shleifer, "Overreaction in macroeconomic expectations," *American Economic Review*, 2020, *110* (9), 2748–82.

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

Cowan, Nelson, *Attention and memory: An integrated framework*, Oxford University Press, 1998.

— , "The magical mystery four: How is working memory capacity limited, and why?," *Current Directions in Psychological Science*, 2010, *19* (1), 51–57.

— , *Working Memory Capacity*, Psychology press, 2012.

— , "The many faces of working memory and short-term storage," *Psychonomic Bulletin & Review*, 2017, *24* (4), 1158–1170.

—, "Working memory: The information you are now thinking of," in John H. Byrne, ed., *Learning and Memory: A Comprehensive Reference (Second Edition)*, second edition ed., Oxford: Academic Press, 2017, pp. 147–161.

De Bondt, Werner, "Betting on trends: Intuitive forecasts of financial risk and return," *International Journal of Forecasting*, 1993, *9* (3), 355–371.

Dwyer, Gerald P, Arlington W Williams, Raymond C Battalio, and Timothy I Mason, "Tests of rational expectations in a stark setting," *Economic Journal*, 1993, *103* (418), 586–601.

Evans, Jonathan BT, "Dual-processing accounts of reasoning, judgment, and social cognition," *Annu. Rev. Psychol.*, 2008, *59*, 255–278.

Frydman, Cary and Gideon Nave, "Extrapolative beliefs in perceptual and economic decisions: Evidence of a common mechanism," *Management Science*, 2016, *63* (7), 2340–2352.

Hay, Dennis C, Mary M Smyth, Graham J Hitch, and Neil J Horton, "Serial position effects in short-term visual memory: A SIMPLE explanation?," *Memory & Cognition*, 2007, *35* (1), 176–190.

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

Hitch, Graham J, Yanmei Hu, Richard J Allen, and Alan D Baddeley, "Competition for the focus of attention in visual working memory: Perceptual recency versus executive control," *Annals of the New York Academy of Sciences*, 2018, *1424* (1), 64–75.

Hu, Yanmei, Richard J Allen, Alan D Baddeley, and Graham J Hitch, "Executive control of stimulus-driven and goal-directed attention in visual working memory," *Attention, Perception, & Psychophysics*, 2016, *78* (7), 2164–2175.

Ilut, Cosmin and Rosen Valchev, "Economic agents as imperfect problem solvers," *The Quarterly Journal of Economics*, 2023, *138* (1), 316–362.

Khaw, Mel Win, Ziang Li, and Michael Woodford, "Cognitive imprecision and small-stakes risk aversion," *Review of Economic Studies*, 2018.

Mankiw, Gregory and Ricardo Reis, "Sticky information versus sticky prices: A proposal to replace the New Keynesian Philips Curve," *Quarterly Journal of Economics*, 2002, *117* (4), 1295–1328.

Phillips, WA and DFM Christie, "Components of visual memory," *Quarterly Journal of Experimental Psychology*, 1977, *29* (1), 117–133.

Poole, Bradley J and Michael J Kane, "Working-memory capacity predicts the executive control of visual search among distractors: The influences of sustained and selective attention," *Quarterly Journal of Experimental Psychology*, 2009, *62* (7), 1430–1454.

Reimers, Stian and Nigel Harvey, "Sensitivity to autocorrelation in judgmental time series forecasting," *International Journal of Forecasting*, 2011, *27* (4), 1196–1214.

Schmalensee, Richard, "An experimental study of expectation formation," *Econometrica*, 1976, pp. 17–41.

Spillers, Gregory, Gene Brewer, and Nash Unsworth, "Working memory and information processing," *Encyclopedia of the Sciences of Learning, New York: Springer*, 2012, pp. 3474–3476.

Sternberg, Saul, "High-speed scanning in human memory," *Science*, 1966, *153* (3736), 652–654.

Unsworth, Nash and Gregory J Spillers, "Working memory capacity: Attention control, secondary memory, or both? A direct test of the dual-component model," *Journal of Memory and Language*, 2010, *62* (4), 392–406.

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