A Comprehensive Empirical Evaluation of Biases in Expectation Formation

Kenneth Eva Fabian Winkler*

June 14, 2023

Abstract

We revisit predictability of forecast errors in macroeconomic survey data, which is often taken as evidence of behavioral biases at odds with rational expectations. We argue that to reject rational expectations, one must be able to predict forecast errors out of sample. However, the regressions used in the literature often perform poorly out of sample. The models seem unstable and could not have helped to improve forecasts with access only to available information. We do find some notable exceptions to this finding, in particular mean bias in interest rate forecasts, that survive our out-of-sample tests. Our findings help narrow down the set of biases that merit closer attention of researchers in behavioral macroeconomics.

JEL: C53, D84, E37

Keywords: Behavioral Bias, Forecasting, Out-of-sample prediction, Rational expectations, Survey Data

*Federal Reserve Board, 20th St and Constitution Ave NW, Washington DC 20551, email: kenneth.j.eva@frb.gov and fabian.winkler@frb.gov. We thank Andrew Chen, Sai Ma, Andrew Patton, and participants at the 2023 AEA Annual Meeting and the 2022 Computing in Finance and Economics conference for helpful comments. The views expressed in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System or any other person associated with the Federal Reserve System.

1 Introduction

Ever since the the theory of rational expectations has entered economics several decades ago, economists have debated the empirical question whether people's expectations are in fact rational. For the purpose of this paper, we say expectations are rational when "outcomes do not differ systematically (i.e., regularly or predictably) from what people expected them to be" (Sargent, 2007). The empirical question then becomes: Are forecast errors predictable using available information, and if so, how?

In this paper, we want to contribute to this debate by drawing a parallel to the question of return predictability in finance. It is easy to show that (aggregate) stock market returns are predictable by regressing them on a readily observed variable like the price/dividend ratio (Fama and French, 1988). This observation led to the development of theoretical models and investment recommendations consistent with the observed patterns of predictability. But the performance of the regressions had been evaluated almost exclusively *in sample* (IS). In an influential study, Welch and Goyal (2007) argued that the in-sample performance is not a useful definition of predictability, because predictions are made using future data that is not available to an investor predicting returns in real time. They instead evaluated the predictive performance *out of sample* (OOS), and found that the predictability previously documented in the literature all but disappeared.

Similarly, it is relatively easy to show that forecast errors are predictable by regressing them on a readily observed variable like forecast revisions. The predictable portion of the forecast error is then called a bias. Many such biases have been documented using different data sets on expectations. A host of labels such as "overreaction", "underreaction", "extrapolation", or "stickiness" have been proposed in an attempt to interpret them. A sizable literature has been devoted to the development of theoretical models of expectation formation to match the empirical findings, examine the propagation of macroeconomic shocks through expectations, and even to make recommendations for the conduct of policy.[1] But again, predictive performance is almost exclusively evalu-

[1] Some examples in this literature are Angeletos et al. (2018); Winkler (2020); Pfäuti and Seyrich (2022); Bhandari et al. (2022).

ated *in sample*.

In our view, the in-sample approach for detecting biases in expectations suffers from the same problem as that for detecting predictability of returns. The in-sample performance is not a useful yardstick to reject the null of rational expectations, because the regression makes use of future data that could not possibly have been available to the agent who formed the expectation. Armed with the benefit of hindsight that an in-sample test provides, it is easy to say that someone's expectations were biased. A higher bar is to demonstrate that more accurate predictions are possible in real time. This is what we set out to do in this paper by evaluating the predictive performance *out of sample*.

We comprehensively re-examine the empirical evidence on the predictability of forecast errors in survey data documented in the literature as of 2022. We find that many so-called biases are unstable or spurious. By and large, the models have poor OOS performance. At best, they predict forecast errors of some variables in some time periods without a clear pattern. Additionally, many regressions are not significant even in-sample. Our evidence suggests that many empirical models of behavioral expectation formation would not have helped to improve their forecasts. In fact, trying to use the models to correct for expectational biases would have led to larger forecast errors.

Importantly, we also find exceptions to this finding. Some biases are remarkably stable out of sample. We find robust evidence of a mean bias in professional forecasts of bond yields across the maturity curve. These patterns could be consistent with rigid priors about the low-frequency behavior of interest rates (Farmer et al., 2021). Also, there is strong evidence for what we call forecast combination bias—the fact that individual forecasts exhibit excess dispersion around the cross-sectional mean. One candidate explanation for this pattern is strategic interaction between forecasters (Gemmi and Valchev, 2021).

Our paper is simple, and we hope that the simplicity of our approach strengthens the credibility of our evidence. We do not view our contribution as a judgment in favor of, or against, rational expectations. Rather, we want to carefully examine which of the previously documented biases are actually useful to improve forecasts in prac-

tice, much like Welch and Goyal (2007) asked which asset pricing models could actually predict returns in practice. Our results can be seen as providing a selection criterion that helps narrow down the set of biases that merit closer investigation in behavioral macroeconomics. Researchers interested in theoretical models of expectation formation may want to focus on replicating those biases that survive our out-of-sample tests.

Why is it that some of the biases that were found to be statistically significant in previous studies do not work in our OOS tests? One candidate explanation is that our tests suffer from low power. It is true that, if a researcher is confident in a well-specified, stable underlying model, an OOS test will always have lower power than an IS test. But this confidence is rarely warranted in practice. Even a solid finding that one's expectations were biased in the past is of limited use if it does not lead to a prescription of how to improve one's expectations in the future.

From the evidence in this paper, we conclude that a more likely explanation for weak OOS performance is that the models are not stable over time. Some of the biases documented in the literature may be time- or state-dependent.[2] As a result, OOS prediction is difficult because one not only has to estimate the bias, but also how it will change in the future. This argument has also been made in the finance literature: Structural changes in return prediction models likely explain the disconnect between IS and OOS predictability (Lettau and Van Nieuwerburgh, 2007). However, we also document some biases that do seem structurally stable: Mean bias in professional forecasts of interest rates, as well as excess dispersion of forecasts around the cross-sectional mean. We think these biases merit closer attention of researchers in the field.

The only recent study we are aware of that takes a rigorous OOS perspective when assessing expectational biases is Bianchi et al. (2022). They show that a sophisticated machine-learning algorithm can improve GDP and inflation forecasts of professional forecasters out of sample. They also document that the predictive regressions of Coibion and Gorodnichenko (2015) perform poorly out of sample for GDP and inflation. The scope of our paper is different. Rather than constructing a new measure of bias, we

[2] In support of the idea of state-dependent biases, (Angeletos et al., 2021) construct impulse responses of forecast errors to identified macroeconomic shocks and find that forecast errors exhibit different patterns of predictability depending on the type of the shock and the time since it occurred.

comprehensively revisit the existing evidence on expectational biases through the OOS lens. Our paper is closer in spirit to an older literature which asks whether forecasts from time-series models improve on survey forecasts out of sample (e.g. Pearce, 1987; Bonham and Dacy, 1991) and usually concludes that they do not. We modify this exercise and ask whether the models of bias proposed in the literature are able to improve survey forecasts.

Our paper relates to a number of studies that voice skepticism about biases in expectations based on the existing empirical evidence. Andolfatto et al. (2008) argue that in macroeconomic models with infrequent regime shifts, rational agents that have to learn about the new regimes will make forecast errors that seem predictable in small samples. Hajdini and Kurmann (2022) show that this can even be the case when the regimes are observed by agents. More subtly, Farmer et al. (2021) argue that in-sample predictability in small samples can still be consistent with rational Bayesian updating if agents are unsure about the low-frequency behavior of the time series being forecast and have relatively strong priors on it, because it takes a very long time for the effect of the bias induced by the prior to fade away. Our paper takes an entirely empirical approach and shows that, from an OOS perspective, the evidence on predictability is often weak to begin with. There is also a strand of the literature that argues that survey forecasts are optimal, but just not in a mean squared error sense. This could be the case because forecasters have asymmetric or otherwise non-quadratic loss functions (Elliott et al., 2008) or have to deal with Knightian uncertainty (Bhandari et al., 2022). Our paper offers a complementary argument for the optimality of survey forecasts from an OOS perspective, retaining the standard mean squared error criterion.

The remainder of this paper is structured as follows. Section 2 describes the data, and Section 3 lays out our empirical procedure. Section 4 contains our findings for surveys of professional forecasters while Section 5 contains findings for household surveys. Section 6 discusses the robustness of our findings. Section 7 concludes.

2 Data

We use data from the Survey of Professional Forecasters (SPF), BlueChip Financial Forecasts (BC), the Michigan Surveys of Consumers (Michigan) and the Survey of Consumer Expectations (SCE). The surveys differ in their sample length and coverage of forecast variables. We only evaluate numerical point forecasts.

The SPF is the longest-running quarterly survey of macroeconomic forecasts in the United States, starting in 1968. Since 1990, the survey is run by the Philadelphia Fed. In the middle of each quarter, participants are asked to forecast a wide range of variables for the current quarter and each of the following quarters, up to four quarters out. From this survey, we take the following forecast variables: the GDP deflator, nominal GDP, industrial production; real GDP, consumption, non-residential investment, residential investment, federal government expenditures, as well as state and local government expenditures; housing starts, the unemployment rate, and CPI headline inflation. For all variables except the last two, the forecasts in the data are in levels but we transform them into forecasts of the percent change between the forecast horizon and the quarter preceding the survey date. For CPI inflation, the forecasts are for annualized quarterly inflation rates but we transform them into forecasts of the percent change of the CPI index between the forecast horizon and the quarter preceding the survey. For the unemployment rate, we directly evaluate the level forecasts. We omit other variables in the SPF as they have less than 20 years of data in order to guarantee a reasonable evaluation period for our OOS tests. Our main focus is on a forecast horizon of three quarters, but we also evaluate forecasts from zero quarters ahead (current-quarter nowcasts) to two quarters ahead.

The BlueChip survey is a monthly survey of forecasts mainly of interest rates. Our BlueChip sample starts in 1988. Because the BlueChip forecast horizons are quarterly, we restrict our sample to the months in the middle of every quarter to ensure constant forecast horizons. From BlueChip, we take forecasts of the federal funds rate; Treasury yields at three months, one year, two years, ten years, and 30 years maturity; and Aaa and Baa corporate bond yields. We also construct implicit forecasts of the one year-

three month and ten year-two year term spread, as well as the Baa-Aaa corporate bond spread.

While the respondents to these two surveys are professional forecasters, the Michigan survey and the SCE are monthly household surveys. The Michigan survey starts in 1978 while the SCE starts in 2013. We only use 12-month ahead inflation forecasts, as those are the only quantitative forecasts of traditional macroeconomic data available.

To construct forecast errors, we also need realized data. For some, like interest rates which are market-quoted, this is straightforward. But for others, like GDP, the realized values are subject to considerable revisions. We use vintage data from the real-time data set for macroeconomists provided by the Philadelphia Fed, and use the first available releases of the data.

3 Empirical procedure

3.1 Consensus forecasts

Many tests for rational expectations in the literature use consensus forecasts, i.e. the cross-sectional average of individual forecasts. This average is then treated as the expectation of a hypothetical aggregate forecaster. In our empirical procedure, we consider the null hypothesis that consensus forecast errors are unpredictable. The regression models we evaluate take the form:

$$y_{t+h} - \bar{y}_{t+h|t} = \beta' x_t + u_{t+h} \tag{1}$$

where y_{t+h} is the realization of a variable at time $t+h$, $\bar{y}_{t+h|t}$ is the consensus forecast of y_{t+h} made at time t, and x_t are a set of K potential predictors, the values of which are known at time t. When the consensus forecast is a rational expectation, the forecast error has zero mean and is unpredictable by x_t; that is, $\beta = 0$. If instead, model (1) captures a behavioral bias, then $\beta \neq 0$.

We use the behavioral model to construct a series of bias-corrected forecasts:

$$y^*_{t+h|t} = \bar{y}_{t+h|t} + \hat{\beta}'_t x_t.$$

When we fit the model IS, then $\hat{\beta}_t$ is constant over time and simply equals the OLS coefficients of (1) estimated over the whole sample. When we fit the model OOS, then $\hat{\beta}_t$ are the OLS coefficients estimated using data available up to time t, either using recursive or rolling windows. In the surveys we consider, the end-of-period values of the forecast variables are not known at time t, so that $\hat{\beta}_t$ is estimated using observations through y_{t-1}.

The prediction errors for the rational model and the behavioral model, respectively, are:

$$\begin{aligned} e^R_{t+h} &= y_{t+h} - \bar{y}_{t+h|t} \\ e^B_{t+h} &= y_{t+h} - \bar{y}_{t+h|t} - \hat{\beta}'_t x_t. \end{aligned}$$

Under the null of rationality, the rational model should predict better than the behavioral model, as the latter is just injecting noise into the prediction. Following the literature, we evaluate the accuracy of forecasts using the sum of squared errors (SSE). We divide the sample into a training period and an evaluation period, the latter starting at time t_0, and compute:

$$SSE^m_t = \sum_{s=t_0}^{t} \left(e^m_s\right)^2, m = R, B. \tag{2}$$

Our main statistic of interest is the difference of the SSE of the rational model and the behavioral model:

$$\Delta SSE_t = \frac{SSE^R_t - SSE^B_t}{SSE^R_T}. \tag{3}$$

If the difference is positive, then the behavioral model predicted better in our sample. If it is negative, then the rational model predicted better. We divide this difference by SSE^R_T, the sum of the squared forecast errors over the entire evaluation period, to allow

for an easier interpretation of the magnitudes. ΔSSE_t thus represents the difference in predictive performance of the rational and behavioral model up to time t, expressed as a fraction of the total sum of squared original forecast errors in the data. A value of, say, $\Delta SSE_T = 0.1$ means that the rational model produces squared forecast errors that are 10 percent larger over the sample than the behavioral model; in other words, correcting for bias using the behavioral model reduces squared forecast errors by 10 percent.[3]

Although we could use statistical tests for equal forecast accuracy of nested models (e.g. Clark and West, 2007), we obtain critical values for ΔSSE_t (both IS and OOS) directly using a bootstrap, as we use relatively small samples, potentially serially correlated independent variables, and overlapping observations whenever $h > 1$.

Our bootstrap follows Welch and Goyal (2007), adapted to the particular structure of overlapping forecasts, and imposes the null of no predictability of forecast errors. The data-generating process for our bootstrap is:

$$u_t = \sum_{s=0}^{h} \theta_s \epsilon_{t-s} \tag{4}$$

$$x_{k,t} = \sum_{s=0}^{p} \phi_s x_{k,t-1} + \sum_{s=0}^{q} \psi_s \eta_{k,t-s},\ k = 1, \ldots, K \tag{5}$$

We model forecast errors as a MA(h) process and regressands as ARMA(p,q) processes, and estimate the parameters by maximum likelihood using the full sample of observations. The choice of p and q depends on the particular model. The joint residuals $(\epsilon_t, \eta_{1t}, \ldots \eta_{Kt})$ are stored for sampling. Joint sampling preserves the correlation structure between the variables. We then generate 10,000 bootstrapped time series by drawing with replacement from the residuals. The initial observation x_{-1} is selected by picking one date from the actual data at random. For each draw, we compute ΔSSE_t and use the resulting distribution to compute critical values. We use one-sided critical values because we are only interested in the whether the behavioral model predicts better than the null.

[3] Note that, for IS regressions, ΔSSE_T is not the same as R^2 because we only sum the squared errors over the evaluation period starting in t_0, and therefore can also be negative. In Welch and Goyal (2007), it is named "IS for OOS $\bar{R}^2$".

To see that it is appropriate to model forecast errors as MA(h), consider that, under the null,

$$u_t = y_{t+h} - E\left[y_{t+h} \mid \mathcal{F}_t\right] = \sum_{s=0}^{h} \left(E\left[y_{t+h} \mid \mathcal{F}_{t+s+1}\right] - E\left[y_{t+h} \mid \mathcal{F}_{t+s}\right]\right) \tag{6}$$

where $\mathcal{F}_t$ is the information set at time t; in particular, y_t and x_t are part of this information set. Each of the forecast revisions in the sum in (4) is uncorrelated with the others, because rational forecast revisions are martingale differences. Also, u_t is uncorrelated with its own lags at lag length greater than h. As long as u_t is also covariance-stationary, then u_t is therefore an MA(h) process. Moreover, x_t is uncorrelated with all forecast revisions that occur after time t.

For our OOS tests, we choose an initial estimation window t_0 of 40 periods, after which we begin the OOS forecasts, and restrict ourselves to forecasts for which at least 80 periods of data are available. Any choice of the window length is necessarily ad-hoc, but our results are robust to this choice, as can be seen in our graphical analysis.

We mainly report results using a recursive window regression to estimate $\hat{\beta}_t$ for our OOS forecasts. It could be argued that this choice makes it harder to produce forecast improvements if the true model coefficients are time-varying, as has been suggested for example by Coibion and Gorodnichenko (2015). However, time variation also makes it harder to predict forecast errors in real time even if properly accounted for, because past data now contain less information about future biases. In Section 6.1, we find that using rolling window regressions does little to improve the predictive performance of the behavioral models.

3.2 Individual forecasts

Some of the literature conducts rationality tests directly on individual forecasts. Models of biased expectations at the individual level take the form:

$$y_{t+h} - y_{t+h|it} = \beta' x_{it} + u_{it+h} \tag{7}$$

where $y_{t+h|it}$ is the forecast of y_{t+h} made at time t by individual i, and x_{it} are one or more potential predictors of forecast errors, the values of which are known to individual i at time t.

Again, our null hypothesis is that the individual expectations are rational so that $\beta = 0$, while a behavioral model posits $\beta \neq 0$. As before, $\hat{\beta}_t$ are the OLS coefficients estimated using data available up to time t, and e_{it}^R and e_{it}^B are the prediction errors for the rational model and the behavioral model, respectively. The sum of squared errors (SSE) for the rational and behavioral model, and their difference, are now defined as

$$SSE_t^m = \sum_{s=t_0}^{t} \sum_{i \in \mathcal{I}_s} \left(e_{is}^m\right)^2, m = R, B \tag{8}$$

$$\Delta SSE_t = \frac{SSE_t^B - SSE_t^R}{SSE_T^R}. \tag{9}$$

where $\mathcal{I}_s$ is the subset of individuals for which forecasts are available at time s.

For the bootstrap, we model the individual forecast errors and regressors analogously to the consensus level:

$$u_{it+h} = \sum_{s=0}^{h-1} \theta_s \epsilon_{i,t+h-s} \tag{10}$$

$$x_{ik,t+h} = \sum_{s=0}^{p} \phi_s x_{ik,t-1} + \sum_{s=0}^{q} \psi_s \eta_{ki,t-s},\ k = 1, \ldots, K \tag{11}$$

where the choice of p and q depends on the particular model. We estimate the parameters using maximum likelihood, now pooling parameter estimates across forecasters, and store the estimated residuals for sampling. When we sample the residuals, we preserve cross-sectional correlations and missing values in the following way. We first sample $T+1$ time indices randomly with replacement from $\{0, \ldots, T\}$. For each time period in the bootstrapped sample, we sample with replacement from the individual residuals only within the corresponding sampled time index. Where possible, we jointly sample residuals from u_{it} and x_{it}, preserving correlation at the individual level. The panel that we obtain is balanced, while the original data have many missing values. In the last step, we therefore replace simulated data with missing values wherever the original data has

missing values. We repeat this simulation 10,000 times.

This bootstrap implies that forecast errors are not only unpredictable at the individual level, but also at the consensus level. The literature has pointed out that even if individual forecasters make rational predictions, average forecasts may still be biased.[4] Our null hypothesis is somewhat stronger than individual rationality and thus easier to reject. If we fail to reject our null for the behavioral models because of weak OOS performance, then the same will hold true for a weaker null in which consensus forecast errors can exhibit some predictability.

4 Results for Professional Forecasters

4.1 Consensus forecast revisions

We first present detailed results of our tests for the popular model of Coibion and Gorodnichenko (2015), which aims to predict consensus forecast errors with forecast revisions. The model posits that forecast errors are predictable using forecast revisions:

$$y_{t+h} - \bar{y}_{t+h|t} = \beta \left(\bar{y}_{t+h|t} - \bar{y}_{t+h|t-1} \right)' + u_{t+h} \tag{12}$$

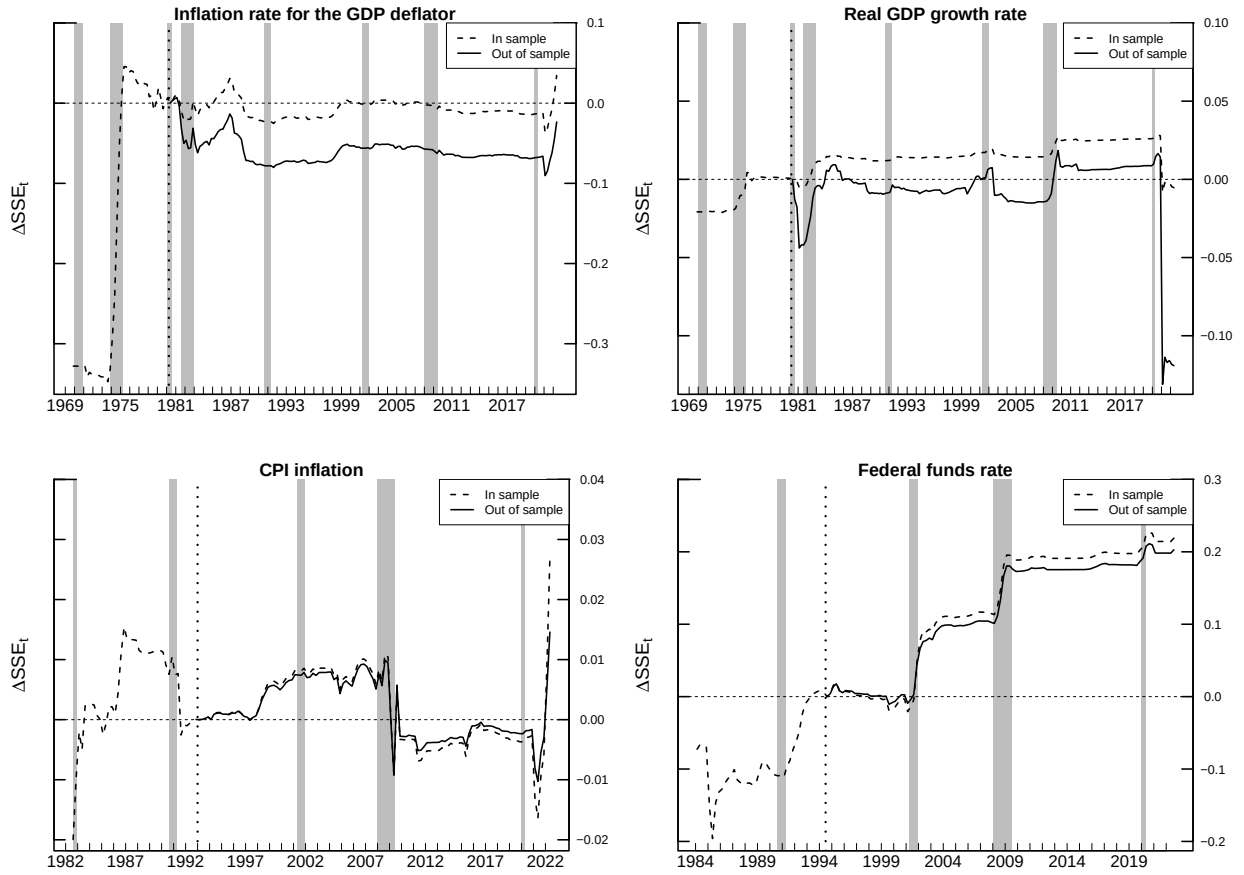
The in-sample coefficient in these regressions is typically positive, which can be interpreted as underreaction: When forecasters revise their expectations upwards, they still make a positive forecast error and thus should have revised more.[5] For this model, we set $p = q = 0$ in (5) as rational forecast revisions are uncorrelated. Coibion and Gorodnichenko (2015) focus mainly on inflation expectations, but also apply the model to a wide range of other variables.

We will first present the results of this model graphically, by plotting the series ΔSSE_t

[4] This is the case, for example, in the noisy information model of Coibion and Gorodnichenko (2015). In that model, the average forecast revision that appears on the right-hand side of the bias regression is never observed by individuals.

[5] The empirical model estimated by Coibion and Gorodnichenko additionally has a constant term. We omit the constant here to give the models the best chance to fit the data OOS because including a constant makes it considerably harder to reject the null, i.e. reduces the power of our OOS tests. Since forecast errors and forecast revisions have zero mean under the null, setting the constant to zero is a reasonable economic prior. Later, we will test for mean bias in forecast errors separately.

Figure 1: Prediction of consensus forecast errors with revisions.



Note: Dashed and solid lines represent cumulative squared errors ΔSSE_t for the in-sample regression and the out-of-sample regression, respectively. An increase in a line indicates better performance of the behavioral model; a decrease in a line indicates better performance of the rational model. Dotted vertical lines mark the end of the training period and the beginning of the evaluation period. Shaded areas represent NBER recessions.

over time. The ΔSSE_t statistic represents the difference of the squared original forecast errors and the squared model prediction errors (either IS or OOS) cumulated up to time t. Thus, whenever a line increases, the behavioral model predicted better; whenever it decreases, the rational model predicted better. The endpoint of the lines represent the improvement in the mean squared forecast error. For example, a value of 0.1 at the end of our sample would imply that a forecaster who used the behavioral model to correct their predictions would have made forecasts with a ten percent lower mean squared error. Figure 1 shows the evolution of ΔSSE_t over time for four variables that represent typical outcomes of our tests.

The top left panel shows results for forecasts of inflation as measured by the GDP deflator, going back to 1968. The IS fit, represented by the dashed line, is surprisingly weak. The behavioral model fits the data better over the entire sample by construction of the OLS estimator. But this fit is largely achieved during a short period around 1975 and to a smaller extent after the Covid-19 recession of 2020. During most of the time, the IS line is flat, indicating that the behavioral model did not fit the data any better than the rational model. The OOS fit, represented by the solid line, is poor. The solid line is below zero for almost the entire evaluation period (to the right of the vertical dotted line). This means that, if a forecaster had used the Coibion and Gorodnichenko (2015) model to improve forecasts in real time, they would have made *larger* forecast errors than if they had treated the original forecasts as optimal predictions.

The top right panel shows the same results for real GDP growth. The IS fit, represented by the dashed line, is again surprisingly weak. Out of sample, the behavioral model beats the rational model in the period after the financial crisis of 2008, when the solid line moves up markedly. After that, however, the OOS line stayed flat, indicating that the behavioral model had little predictive advantage. And after the Covid-19 recession in 2020, the behavioral model fared terribly. Importantly, this bad performance is not a mechanical result of the large forecast errors realized in 2020. The OOS line would have stayed flat if the behavioral model had made forecast errors that were as large as the unadjusted forecasts. But instead, the behavioral model modified the survey forecasts in the wrong direction, resulting in even larger forecast errors. When the pandemic shock hit the economy, economic activity forecasts revised down and the behavioral model adjusted the forecasts down further, in line with the underreaction of expectations. But subsequently, economic activity rebounded more quickly than implied by the original forecasts, which would be more consistent with overreaction. This is a good example of an unstable predictive relationship.

The model does somewhat better on expectations of CPI inflation, shown in the bottom-left panel. CPI inflation expectations are the main focus of Coibion and Gorodnichenko, and here their model would have led to a 1 percent reduction in the mean squared forecast error at the end of the sample. This improvement is sufficient to reject

the null using our bootstrapped critical values. However, the figure also shows that the gains over the rational model arise largely in 2021 and 2022, when inflation forecasts indeed underreacted to a sharp rise in inflation. Between 2007 and 2020, the bias correction of the model would have made the forecasts worse. To us, this also looks like an unstable model.

Remarkably, the Coibion and Gorodnichenko model does very well on interest rate forecasts. The lower right panel of Figure 1 shows that federal funds rate forecasts could have been substantially improved using this model. What's more, the performance is indicative of a stable prediction model: The OOS line trends up almost through the entire sample. At the end of our sample, a forecaster relying on the behavioral model to correct their forecasts would have reduced the mean squared error of their predictions by an impressive 19 percent.

Our results do not depend materially on the split between the training and the evaluation period, and this can be seen directly from the plots. The IS and OOS lines represent cumulative sums of squared errors, and if we start the evaluation period at a later date, then we can simply start summing the squared errors from that date.

In Table 1, we document the IS and OOS performance of the Coibion and Gorodnichenko model, measured by the ΔSSE_T statistic in (1)–(2), for all variables in the SPF and BlueChip surveys with three-quarter ahead forecast horizons for which at least 20 years of data are available. Stars indicate whether the the null of no predictability is rejected using our bootstrapped critical values for the IS and OOS versions of ΔSSE_T.

For almost all of the macroeconomic variables in the top half of the table, even the IS predictive performance is not high enough to reject the null of rationality using our bootstrap test. Note that the statistic ΔSSE_T is at times negative IS. If we summed over the full sample in (2), the statistic would equal R^2 and would be positive by construction of the OLS estimator, but we only sum squared errors over the evaluation period, that is, excluding the first 40 quarters of the sample. The fact that the IS performance is often not significant during this evaluation period already indicates an unstable predictive relationship.

The OOS performance for the macroeconomic variables is generally negative, inply-

ing that a forecaster who had relied on the Coibion-Gorodnichenko method to remove bias from their forecasts would have been left worse off than with no adjustment at all. There are some notable exceptions to this rule, however. The behavioral model beats the rational model OOS for CPI inflation, industrial production and housing starts. The improvements are significant and reduce the mean squared forecast error by several percentage points. Where the model does really well is on interest rate forecasts, shown in the bottom half of the table. A forecast correction using the behavioral model would have led to a reduction in the mean squared forecast errors by up to 21 percent, which are staggering numbers by forecasting standards.

4.2 Other models based on consensus forecasts

Moving beyond the prominent model of Coibion and Gorodnichenko (2015), we now turn to discuss other widely known models of bias in consensus expectations.

First, we examine a simple model of mean bias, where forecasts are systematically too high or too low. To test for mean bias, consensus forecast errors are regressed only on a constant:

$$y_{t+h} - \bar{y}_{t+h|t} = \beta + u_{t+h} \tag{13}$$

A positive coefficient implies that forecasters always underpredict a variable by a constant amount. Because of its simplicity, we expect this bias to be the easiest to detect.

Next, we test for forecast autocorrelation, whereby forecast errors are predicted with their own lag:

$$y_{t+h} - \bar{y}_{t+h|t} = \beta \left(y_{t-1} - \bar{y}_{t-1|t-h-1} \right) + u_{t+h} \tag{14}$$

A positive coefficient on the lagged forecast error implies that overpredictions tend to be followed by more overpredictions, akin to a momentum effect in asset returns, implying that forecasts are slow to react to incoming information. Note that, in the data, forecasters at time t only know the realizations of the data (and thus their own forecast errors) up to period $t-1$. For the bootstrap, we set $p = 0$, $q = h + 1$.

We also look at the well-known regressions of Mincer and Zarnowitz (1969), in which

Table 1: Prediction of consensus professional forecast errors with revisions.

ΔSSE_T	IS	OOS
Inflation (deflator)	0.034**	-0.023
Inflation (CPI)	0.027**	0.015**
Real GDP	-0.005	-0.119
Industrial Production	0.094***	0.035***
Nominal GDP	0.021**	-0.101
Unemployment rate	0.003	-0.251
Consumption	0.010	-0.031
Non-residential inv.	0.019**	-0.061
Residential inv.	0.032***	-0.026
Federal govt.	-0.002	-0.012
Non-federal govt.	0.002	-0.027
Housing starts	0.132***	0.047***
Federal funds rate	0.219***	0.203***
3-month yield	0.190***	0.181***
6-month yield	0.234***	0.211***
1-year yield	0.220***	0.196***
2-year yield	0.143***	0.112***
10-year yield	0.025*	-0.001
Aaa yield	0.069***	0.067***
Baa yield	0.061**	0.052**
1y-3m spread	-0.002	-0.009
10y-2y spread	0.083***	0.056***
Aaa-Baa spread	-0.004	-0.011

Note: Each row shows cumulative squared errors ΔSSE_T in sample and out of sample. ***, ** and * represent rejection of the null hypothesis of no predictability of forecast errors at the 10, 5, and 1 percent level using bootstrapped critical values. Yield and spread variables are taken from BlueChip, other variables are taken from the SPF.

realized outcomes are regressed on forecasts. An equivalent formulation is to regress forecast errors on forecasts and a constant:

$$y_{t+h} - \bar{y}_{t+h|t} = \beta_0 + \beta_1 \bar{y}_{t+h|t} + u_{t+h} \tag{15}$$

Again, the null hypothesis implies $\beta = 0$. For this model, we set $p = 2$ and $q = 0$ in the bootstrap (5). A positive coefficient on the forecast $\bar{y}_{t+h|t}$ implies that forecasters are too optimistic whenever their forecasts are high, thus capturing a form of extrapolation bias.

Finally, we look at the Nordhaus (1987) test of forecast efficiency. Instead of putting forecast errors on the left-hand side, Nordhaus examined the predictability of forecast revisions. Rational expectations imply that forecast revisions are unpredictable, in addition to forecast errors, because of the law of iterated expectations.[6] The Nordhaus regression model has the form:

$$\bar{y}_{t|t-h} - \bar{y}_{t|t-h-1} = \beta \left(\bar{y}_{t|t-h-1} - \bar{y}_{t|t-h-2} \right) + u_{t+h} \tag{16}$$

This model can be interpreted as a test of the stickiness of forecasts: A positive coefficient on past revisions implies underreaction of forecasts, as the forecasts will be predictably revised in the same direction as the previous revision. For the bootstrap, here we model both the regressor and the regressand as white noise ($p = q = 0$).

The results of our OOS tests for these models are documented in Table 2. For the macroeconomic variables, a similar picture emerges regardless of the model: The IS performance is typically weak and insufficient to reject the null, and the OOS performance is typically negative: A real-time bias correction would have made the forecasts worse. In general, none of the models are able to consistently beat the null of no predictability. There are some exceptions throughout the table. For example, forecast errors of deflator-based inflation and housing starts are predictable OOS using the autocorrelation model. To us, these isolated "wins" are likely to arise by chance, as a byproduct

[6] Strictly speaking, unpredictability of forecast revisions also requires that the information sets are nested over time, so that no information is "forgotten" as the forecasts are revised.

of the large number of tests in this paper. The picture looks a bit better for the Nordhaus model, which only predicts forecast revisions instead of actual forecast errors, and sometimes does well at that. This better performance relative to the other models may reflect that revisions contain less noise than realizations and are therefore more easily predictable. Nevertheless, the quest for a simple, unifying empirical relationship summarizing bias in macroeconomic expectations of professional forecasters seems elusive.

The picture is different for interest rate forecasts, shown in the bottom half of the table. Here, all models are able to improve forecast efficiency OOS for short-term yields. The mean bias model in particular is able to reduce the mean squared forecast error across all interest rates in the table, by as much as 29 percent. Our interpretation of this strong deviation from rationality is that mean bias is driving the performance of the other models, too. The level of interest rates has declined steadily over the past few decades, and while forecasters continually revised their projections also, they consistently underestimated the secular fall in interest rates (Rungcharoenkitkul and Winkler, 2022). As a result, forecast revisions and forecast errors are on average negative, resulting in predictability in the models of Coibion and Gorodnichenko (2015) and Nordhaus (1987); forecast errors are positively correlated; and the Mincer and Zarnowitz (1969) also performs well as it includes a constant that picks up the mean bias.

In sum, none of the models that feature prominently in the literature can consistently improve forecasts of macroeconomic expectations. However, interest rate forecast errors are robustly predictable with many of these models, perhaps related to the persistent underprediction of the secular decline in interest rates.

4.3 Tests based on individual expectations

One can argue that consensus forecasts, which average out the idiosyncrasies of individuals, represent a "best case": If it can be shown that average forecasts are biased, then the individual forecasts must be biased as well. This argument is generally valid as long as the predictor variable is a part of the information set of all individuals. But using consensus forecasts is only an indirect way of testing for biases, because they do

Table 2: Other models of consensus professional forecast errors.

	(1)		(2)		(3)		(4)	
	Mean bias		Autocorrelation		Mincer-Zarnowitz		Nordhaus	
ΔSSE_T	IS	OOS	IS	OOS	IS	OOS	IS	OOS
Inflation (deflator)	-0.025	-0.416	0.128***	0.122***	-0.061*	-0.476	0.094***	0.066***
Inflation (CPI)	-0.008	-0.073	-0.001	-0.040	0.013*	-0.043	0.105***	0.075***
Real GDP	0.006	-0.040	0.029*	0.005	-0.004*	-0.153	0.004	-0.074
Industrial Production	0.055*	0.007	0.006	-0.008	0.082*	0.008	0.055***	-0.007
Nominal GDP	0.025	-0.029	0.018	-0.028	0.022*	-0.052	0.013	-0.061
Unemployment rate	-0.001	-0.029	0.005	-0.032	0.048	-0.041	0.000	-0.126
Consumption	0.012	-0.011	0.052*	-0.172	0.021*	0.010	0.027**	-0.082
Non-residential inv.	-0.004	-0.065	0.000	-0.035	-0.001	-0.112	0.075***	0.011**
Residential inv.	0.002	-0.063	0.075**	-0.007	-0.041	-0.306	0.128***	0.098***
Federal govt.	0.001	-0.062	0.044*	-0.035	0.118*	0.027	-0.004	-0.014
Non-federal govt.	0.025	-0.105	0.056*	0.026	0.082*	-0.103	0.107***	0.084***
Housing starts	0.008	-0.043	0.141***	0.117***	0.027*	-0.100	0.233***	0.205***
Federal funds rate	0.121**	0.061**	0.071*	0.047**	0.134*	-0.007	0.234***	0.222***
3-month yield	0.183***	0.129***	0.113**	0.089***	0.221**	0.089**	0.230***	0.217***
6-month yield	0.211***	0.157***	0.186***	0.129***	0.255**	0.115**	0.259***	0.245***
1-year yield	0.198**	0.135**	0.152*	0.037	0.208*	0.043	0.240***	0.227***
2-year yield	0.212***	0.154***	0.128**	-0.022	0.229**	0.046**	0.187***	0.168***
10-year yield	0.323***	0.295***	0.040	-0.046	0.337***	0.054**	0.068***	0.042***
Aaa yield	0.268***	0.225***	0.044	0.016	0.260**	-0.125	0.101***	0.093***
Baa yield	0.439***	0.402***	0.172**	0.108**	0.442***	0.238***	0.135***	0.097***
1y-3m spread	-0.019	-0.118	0.030	0.011	0.093	-0.148	0.036**	0.017**
10y-2y spread	0.002	-0.061	0.008	-0.022	0.078	-0.057	0.118***	0.090***
Aaa-Baa spread	-0.036	-0.139	0.075	0.071	0.153	0.036	-0.040	-0.129

Note: Each row shows cumulative squared errors ΔSSE_T in sample and out of sample for a number of predictive models of forecast errors. ***, ** and * represent rejection of the null hypothesis of no predictability of forecast errors at the 10, 5, and 1 percent level using bootstrapped critical values. Yield and spread variables are taken from BlueChip, other variables are taken from the SPF.

not represent the forecasts of any one individual. The literature has also documented biases in expectations at the individual level, to which we now turn.

One of the most prominent recent studies examining the rationality of individual forecasts is Bordalo et al. (2020) (BGMS). They run a regression of the form:

$$y_{t+h} - y_{t+h|it} = \beta \left(y_{t+h|it} - y_{t+h|it-1} \right)' + u_{it} \tag{17}$$

They document that the prediction of forecast errors with forecast revisions also works at the individual level, but often with a negative coefficient on the revision. This negative coefficient is interpreted as overreaction of forecasts: When forecasters raise their forecasts, they tend to overpredict.[7]

We further test the autocorrelation, Mincer and Zarnowitz (1969), and Nordhaus (1987) models at the individual level:

$$y_{t+h} - y_{t+h|it} = \beta \left(y_{t-1} - y_{t-1|it-h-1} \right) + u_{it+h} \tag{18}$$

$$y_{t+h} - y_{t+h|it} = \beta_0 + \beta_1 y_{t+h|it} + u_{it+h} \tag{19}$$

$$y_{t|it-h} - y_{t|it-h-1} = \beta \left(y_{t|it-h-1} - y_{t|it-h-2} \right) + u_{it+h}. \tag{20}$$

A variation of the Mincer and Zarnowitz (1969) model has recently been advanced by Kohlhas and Walther (2021), who regress forecast errors on the realized values of a variable. We only include the lagged value of the realization as the current-period value is not part of the information set when the forecasts are made:

$$y_{t+h} - y_{t+h|it} = \beta_0 + \beta_1 y_{t-1} + u_{it+h} \tag{21}$$

For this model, we set $p = 2$ and $q = 0$ in the bootstrap (11).

Finally, we test a model based on forecast combination. It is well documented that combining forecasts from different people and models almost always improves fore-

[7] Like at the consensus level, we exclude a constant from the regression. We also omit fixed effects and pool the regression coefficient. Including individual-specific parameters would make OOS prediction very hard, due to the small number of observations in the individual samples (less than 10 on average in the SPF).

casting performance in practice (see Timmermann, 2006, for a review). Based on this idea, we construct a model of biased expectations as follows:

$$y_{t+h} - y_{t+h|it} = \beta_0 + \beta_1 \left(\bar{y}_{t-1|t-h-1} - y_{t-1|it-h-1} \right)' + u_{it+h}. \tag{22}$$

The variable on the right-hand side is the lagged difference between the consensus forecast and the individual forecast. The timing is important: We do not relate individual forecasts to the consensus forecast in the same period, since this object is not known to the agents at the time they complete the survey. However, past consensus forecasts as well as the agents' own predictions are part of their information set, making this a valid test of rationality. For the bootstrap (11), here we set $p = 1$, $q = 0$.

We subject these models to OOS tests at the level of individual forecasts. If our failure to reject the null of no predictability at the consensus level were a problem of small sample size and estimation noise, we should expect predictive performance to improve at the individual level, where the cross-sectional dimension of the data greatly expands the number of observations.

In Figure 2, we show charts plotting ΔSSE_t of the first and last of these models, using CPI inflation, unemployment rate, and three-month Treasury bill forecasts.

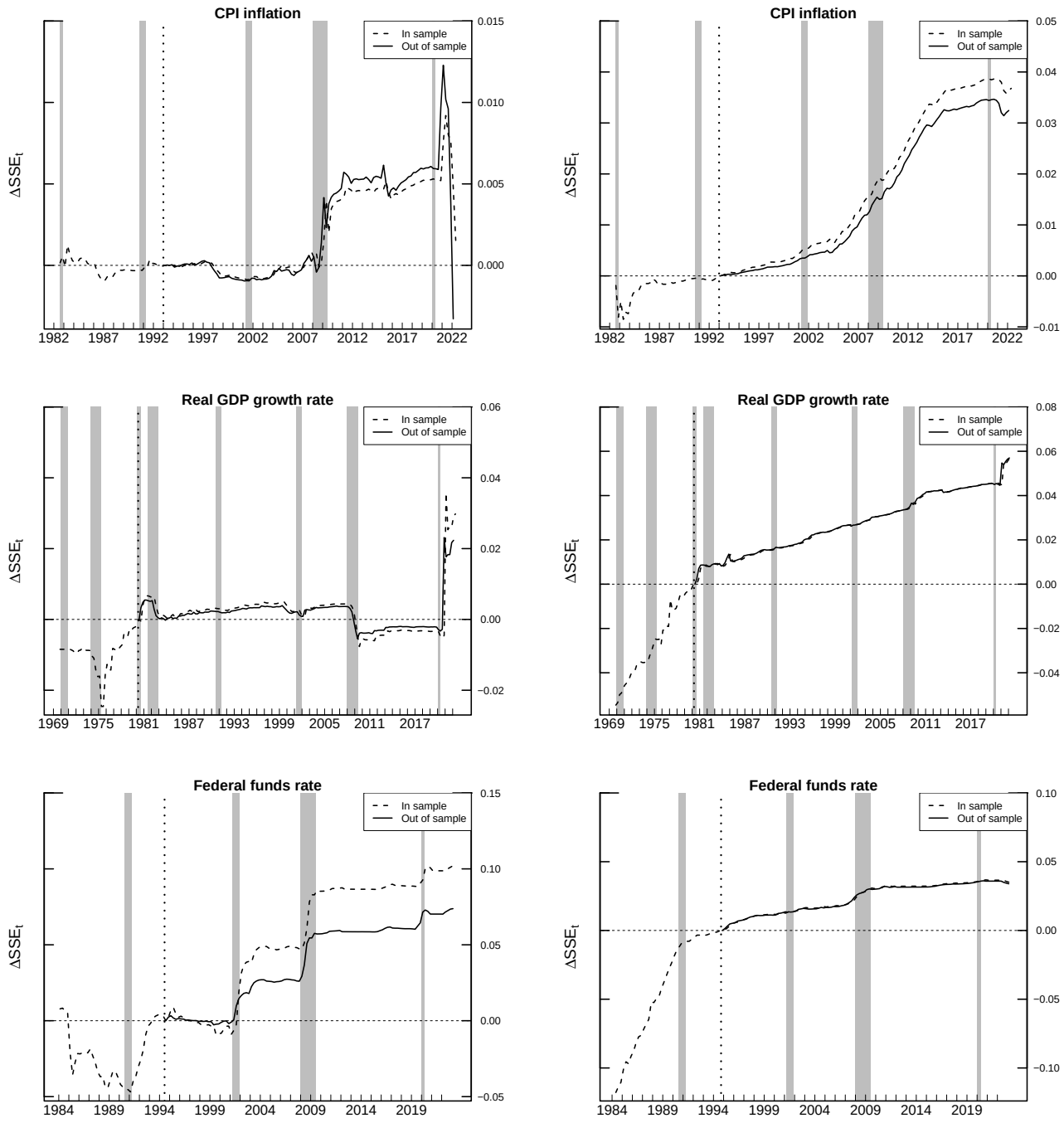
The left panels show the performance of the BGMS model (17). For CPI inflation (top left panel), the model's performance is mixed. It outperforms the rational benchmark between 2008 and 2020, but these gains are erased at the end of the sample. Overall, we cannot reject the null of no predictability. For real GDP growth (middle left panel), the picture is the reverse: Tepid performance during most of the sample, then a big improvement at the end which lifts the OOS performance to levels for which we can reject the null. The estimated coefficient on revisions is negative throughout the sample, which allows the model to fit the overreaction of (aggregate) expectations after the Covid-19 shock in 2020. For federal funds rate forecasts (bottom left panel), the behavioral model outperforms the null of rationality by eight percent, which is less than the Coibion and Gorodnichenko (2015) model at the consensus level but still quite strong.

The right panels of Figure 2 show the performance of the forecast combination bias

Figure 2: Prediction of individual professional forecast errors.

(a) Revisions.

(b) Forecast combination.



Note: Dashed and solid lines represent cumulative squared errors ΔSSE_t for the in-sample regression and the out-of-sample regression, respectively. Dotted vertical lines mark the end of the training period and the beginning of the evaluation period. An increase in a line indicates better performance of the behavioral model; a decrease in a line indicates better performance of the rational model. Shaded areas represent NBER recessions.

model (22). We judge the performance of this model to be remarkable. It consistently performs well across all variables in our data set. For CPI inflation (top right panel), the reduction in the mean squared forecast error is about three percent. Moreover, the null of no predictability consistently is rejected consistently over time: The black line in the chart steadily trends upward, never moving down appreciably. This is a beautiful example of a stable predictive relationship. Indeed, the estimated coefficients $\hat{\beta}_t$ in the OOS prediction also remain stable over the sample. A similar picture emerges for real GDP forecasts (middle right panel): Here, too, the forecast combination bias model outperforms the null consistently, and the performance is roughly double that of the BGMS model. For interest rate forecasts, this model achieves a reduction in the mean squared forecast error of close to five percent. What is remarkable is that the gains in predictive performance are accumulated steadily and robustly over time.

Results for all variables are shown in Table 3. Starting in Column (1), the BGMS model predicting forecast errors with revisions manages to achieve significant performance gains for a number of variables. This model seems to outperform the null of rationality in some areas, but not in others. Overall, we think that it does not represent a bias that is universal in professional forecasts.

Table 3: Prediction of individual professional forecast errors.

	(1)	(2)	(3)	(4)	(5)	(6)
ΔSSE_T, OOS	BGMS	Autocorrelation	Mincer-Zarnovitz	Nordhaus	Kohlhas-Walther	Forecast combination
Inflation (deflator)	0.007***	0.114***	-0.513	0.007***	-0.323	0.141***
Inflation (CPI)	-0.003	-0.036	0.043**	-0.009	-0.036	0.034***
Real GDP	0.022***	-0.029	0.042**	0.004***	0.013	0.060***
Industrial Production	-0.010	-0.016	0.036**	-0.010	0.002	0.049***
Nominal GDP	0.011***	-0.060	-0.071	0.002**	-0.061	0.061***
Unemployment rate	-0.114	-0.048	0.010	-0.065	0.024	0.013***
Consumption	0.058***	-0.166	-0.003	0.003**	-0.090	0.036***
Non-residential inv.	-0.018	-0.035	-0.055	-0.007	-0.102	0.064***
Residential inv.	-0.019	0.046***	-0.037	-0.021	-0.092	0.108***
Federal govt.	0.082***	0.063***	-0.009	-0.014	-0.027	0.148***
Non-federal govt.	0.123***	0.039***	0.126***	-0.007	-0.048	0.206***
Housing starts	0.004	0.208***	-0.099	-0.009	-0.054	0.108***
Federal funds rate	0.080***	0.075***	-0.070	0.083***	0.035**	0.035***
3-month yield	0.071***	0.132***	0.025**	0.063***	0.123***	0.043***
6-month yield	0.108***	0.145***	0.091***	0.089***	0.132***	0.031***
1-year yield	0.085***	0.070***	-0.007	0.071***	0.039**	0.041***
2-year yield	0.040***	0.032***	-0.015	0.032***	0.04**	0.044***
10-year yield	-0.003	-0.016	0.008	-0.012	0.099***	0.069***
Aaa yield	0.000	0.005	-0.181	-0.003	-0.160	0.068***
Baa yield	0.001	0.197***	0.012	-0.043	0.248***	0.130***
1y-3m spread	0.084***	-0.025	0.232***	0.108***	-0.233	0.087***
10y-2y spread	-0.002	-0.004	0.029**	-0.006	0.002	0.053***
Aaa-Baa spread	0.008	-0.084	0.513***	-0.069	0.066	0.067***

Note: Each row shows cumulative squared errors ΔSSE_T in sample and out of sample for a number of predictive models of forecast errors. ***, ** and * represent rejection of the null hypothesis of no predictability of forecast errors at the 10, 5, and 1 percent level using bootstrapped critical values.

Columns (2) through (5) show the performance of the autoregressive model, the Mincer-Zarnovitz model, the Nordhaus model, and the Kohlhas-Walther model. Among these models, the autoregressive model in Column (2) performs the best across the macroeconomic variables, so that there is some evidence that professional forecast errors are persistent at the individual level, though not universally so. For interest rates, it is the Nordhaus model in Column (4) that performs best, and with a similar magnitude as the BGMS model. To us, these findings are consistent with inefficient, slowly mean-reverting deviations of individual forecasts from the average. If individual revisions reflect such deviations, then positive revisions negatively predict forecast errors and future revisions, and forecast errors are autocorrelated. Such inefficient forecast dispersion could arise because forecasters respond to strategic diversification incentives by deviating from optimal forecasts (Gemmi and Valchev, 2021).

The forecast combination bias model in Column (6) precisely represents such inefficient deviations, as it predicts that forecasters that are optimistic relative to the average last period will be too optimistic. This model shows significant predictive gains *across all variables* in our data set. We judge this to be a remarkable achievement. It is worth noting that this result is stronger than the well-known fact that the consensus forecast is more efficient than individual forecasts, because the behavioral model is based off the lagged value, rather than the current value, of the consensus forecast. Combined with the earlier observation that the performance gains accrue steadily over time, we conclude that inefficient dispersion of individual forecasts is the most robust and stable departure from rational expectations in surveys of professional forecasters.

5 Results for Households

So far, we have focused on expectations of professional forecasters. These individuals are usually well-educated specialists employed by financial institutions who expend a great amount of time and resources forming their expectations. Our failure to reject the null of rationality with our OOS tests may in fact indicate that the expectations of these individuals are quite close to rational. However, there also exist surveys of less

sophisticated forecasters, particularly of households. We expect that households form less accurate expectations, and that it should be easier to reject the null of rational expectations.

The two main surveys of American households that elicit macroeconomic expectations are the Michigan survey and the Survey of Consumer Finances (SCE). There are several differences in methodology between these two surveys, but most importantly, the SCE starts in 2013 while the Michigan survey goes back to 1978. We restrict ourselves to 12-month ahead inflation expectations ($h = 12$), as other expectations either have limited coverage or only have categorical response variables. We define realized inflation as headline CPI inflation.

We first aggregate the individual forecasts using both the average and the median, since there are meaningful differences between the two for households. At the aggregated level, we test the models (12)–(15). For the Coibion-Gorodnichenko model (12), we use the month-over-month difference in consecutive 12-month ahead inflation expectations as a proxy for forecast revisions due to data limitations. At the disaggregated individual level, we test a panel version of the mean bias model (13), as well as the Mincer-Zarnovitz model (19) and the forecast combination model (22). For the forecast combination test, we use the difference of the current forecast and last period's consensus forecast to proxy for past disagreement, again due to data limitations. The bootstrap parameters are set in the same way as for the professional forecaster data, except for the forecast errors themselves. Fitting an MA(13) process, which would be natural under the null, is infeasible due to data limitations. Instead, we fit an MA(3) process.

Table 4 summarizes the results of our tests for households.

Column (1) shows the results from the simple mean bias model. In sample, it is relatively easy to detect a (positive) mean bias in household inflation expectations.[8] But out of sample, this mean bias is difficult to exploit because it is hard to estimate its magnitude in real time. The mean bias model still significantly improves Michigan average forecast errors OOS, but fails badly for the shorter sample in the SCE. The dis-

[8]The bias is more pronounced for the average compared to the median, as the distribution of individual inflation forecasts is skewed to the upside.

Table 4: Prediction of household inflation forecast errors.

	(1)		(2)		(3)		(4)		(5)	
	Mean bias		Revisions		Autocorrelation		Mincer-Zarnovitz		Forecast combination	
ΔSSE_T	IS	OOS	IS	OOS	IS	OOS	IS	OOS	IS	OOS
Michigan avg.	0.239***	0.028**	0.003*	0.002	0.086**	0.063***	0.181***	0.008**	–	–
Michigan median	0.012	-0.226	0.001	-0.005	-0.001	-0.043	-0.034	-0.252	–	–
Michigan ind.	0.024***	-0.014	–	–	–	–	0.875***	0.617***	0.896***	0.889***
SCE avg.	0.250**	-0.104	0.000	-0.002	0.178*	0.032	0.355**	-1.491	–	–
SCE median	-0.115	-0.562	0.203***	0.050**	-0.013	-0.061	0.184	-2.139	–	–
SCE ind.	-0.034	-0.113	–	–	–	–	0.742***	0.632***	0.819***	0.799***

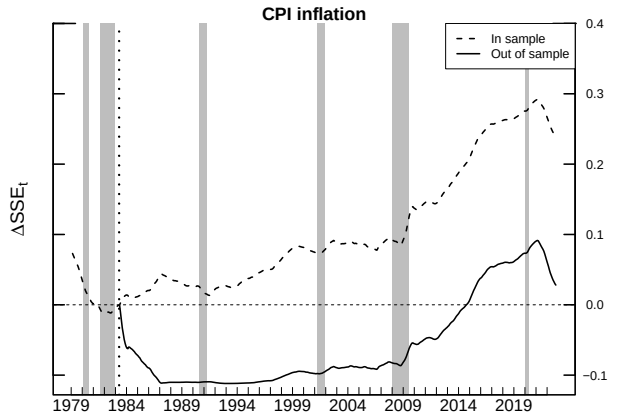
Note: Each row shows cumulative squared errors ΔSSE_T in sample and out of sample for a number of predictive models of forecast errors. ***, ** and * represent rejection of the null hypothesis of no predictability of forecast errors at the 10, 5, and 1 percent level using bootstrapped critical values. For the bootstrap on SCE data, we restrict the lag length of the MA process of u_{it} to 3.

crepancy between the two surveys can be attributed to their different sample windows. To see this, we show the evolution of OOS performance of the mean bias model in the Michigan survey over time in Figure 3. The left panel shows the predictive performance for the average expectation. Starting around 1990, the mean bias model made steady gains as average inflation expectations stayed stubbornly above actual inflation for two decades. However, after 2020, inflation soared much faster than inflation expectations, defying the predictions of a positive mean bias. Over the length of the Michigan sample, the overall predictive performance of the mean bias model is good enough so that we assign statistical significance using our bootstrapped critical values. But in the more limited sample of the SCE (not shown in the figure), the period of high inflation starting in 2021 takes a much greater share of the sample, which explains why the mean bias model underperforms the null of rational expectations in Table 4. Also, the right panel of Figure 3 illustrates that median inflation expectations tracked actual inflation much more closely than average expectations, resulting in a mean bias that averages near zero in sample and dismal OOS performance of the mean bias model applied to median household expectations.

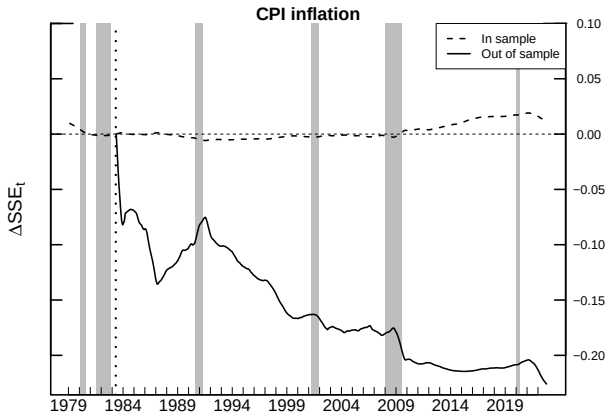
Column (2) of Table 4 shows that using the Coibion and Gorodnichenko (2015) model of regressing forecast errors on revisions does not lead to any significant improvements in Michigan forecast errors, either in sample or out of sample. This model is quite un-

Figure 3: Prediction of Michigan inflation expectations: mean bias model.

(a) Average expectations.



(b) Median expectations.



Note: Dashed and solid lines represent cumulative squared errors ΔSSE_t for the in-sample regression and the out-of-sample regression, respectively. Dotted vertical lines mark the end of the training period and the beginning of the evaluation period. An increase in a line indicates better performance of the behavioral model; a decrease in a line indicates better performance of the rational model. Shaded areas represent NBER recessions.

stable and the estimated coefficient is sometimes positive, sometimes negative. The model does work well for SCE medians, but the entire forecasting performance in that specification is generated in 2022, the last year of the sample. During that year, median forecast errors and median revisions were positive, consistent with underreaction of expectations. By contrast, the model is not able to improve on SCE average forecasts because the average expectation—consistently higher than the median—was remarkably close to actual inflation in 2022.

Column (3) shows the performance of the autoregressive model. Like the mean bias model, this model only works well for Michigan average expectations, but not for SCE averages, or medians in either survey. This pattern is an expression of the same positive mean bias in household average inflation expectations in the last two decades discussed above, because mean bias implies autocorrelation of forecast errors.[9]

Column (4) shows the Mincer-Zarnovitz model. At the mean or median level, this model fares similarly to the mean bias model due to the inclusion of a constant term in

[9]Note that we omit a constant in the autoregressive model, so that mean bias implies a positive coefficient on lagged forecast errors in that model.

that model (in fact, the coefficient β_1 on forecasts in (15) is stable around one). What is most noticeable, however, is the stunning performance of this model at the individual level: More than 60 percent of the mean squared forecast error is predictable OOS using this model in both surveys. The model consistently predicts $\beta_1 \approx -1$ in Equation (19), which means that the individual forecasts are treated entirely as noise, and the improved forecast of that model is just the constant β_0. This behavior is a consequence of the fact that the dispersion in individual household inflation expectations dwarfs the variation in mean forecasts, and so it is best to disregard the individual variation in forecasts.

Excess dispersion of forecasts also explains why the forecast combination model in Column (5) performs well. More than 80 percent of the individual mean squared forecast error can be predicted using this model. The coefficient on the lagged difference of an individual's forecast and the consensus is close to one and stable over time. The fact that the lagged difference predicts individual forecast errors so well points to the persistence in household inflation forecasts.

Summing up, household inflation forecasts do seem much more biased than those of professional forecasters. Their deviation from optimal forecasts occurs mostly at the individual level, where inflation forecasts display an excessive degree of dispersion. At the consensus level, there is some evidence of mean bias, although this bias is not stable over time. Instead, household's mean bias in inflation expectations appears to be time-varying.

6 Robustness

6.1 Rolling window regression

A central theme that emerges from our analysis is that the behavioral models often seem unstable over time. Indeed, in our OOS regressions, many of the estimated model coefficients display sizable variation over time. Biases can be time- or state-dependent. If this is the case, real-time prediction of forecast errors is inherently more difficult, as

one now has to not only estimate past bias, but also predict how the bias will change in the future. This may explain the disconnect between IS and OOS predictability that we have often observed in our analysis so far. This explanation has also been put forward in the finance literature (Lettau and Van Nieuwerburgh, 2007). Traditional tests that have a constant parameter hypothesis, such as the ones we have used so far, are then misspecified.

One way to take time variation into account is to run our regressions with a rolling window instead of a recursive window. In keeping with the simplicity of our empirical approach, we choose rolling windows over a more sophisticated time-varying parameter regression. We keep the initial training sample period of 40 quarters, but now also use this as the size of a rolling window used to estimate the real-time OOS coefficients for prediction. Generally, we find that using rolling window regression does not help to predict forecast errors: The increased estimation noise from smaller window sizes outweighs any gain from capturing time variation in the true model coefficients. Because of the increase in estimation noise, our bootstrapped critical values for rejecting the null also decrease. The bootstrapped significance levels remain fairly similar to those of our baseline estimation. The online appendix contains detailed results of rolling window regressions.

6.2 Other forecast horizons

Our baseline estimation uses three-quarter ahead horizons, which is used by Coibion and Gorodnichenko (2015) and many other studies in the literature. But we also examine the robustness of our results to the choice of the forecast horizon. As documented in the online appendix, the results do not depend materially on this choice. There is some more OOS predictability at the "nowcast" horizon, i.e. of forecast errors of the current-quarter realizations $h = 0$.

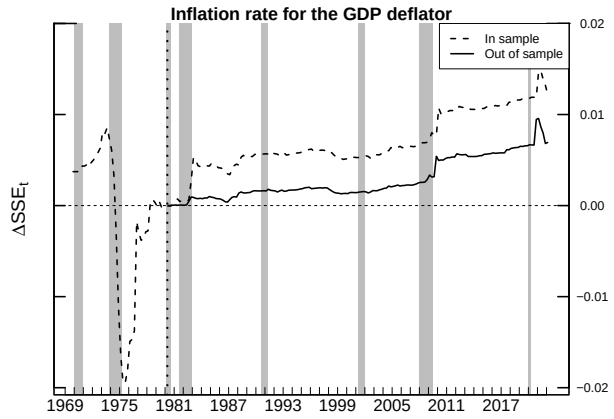
6.3 Adding an intercept

We have omitted an intercept term from the models whenever this can be motivated with an economic prior. For example, in the model of Coibion and Gorodnichenko (2015), forecast errors and forecast revisions both have zero unconditional mean, and yet the former is predictable by the latter, so an intercept is in principle unnecessary for the empirical regression. While having an intercept in an IS regression is common and amounts to nothing more than demeaning the data, in an OOS test this decision can have important consequences. An intercept is another parameter that needs to be estimated in real time, increasing estimation noise. Moreover, in small samples, distinguishing between the contributions of a highly autocorrelated variable and a constant can be challenging.

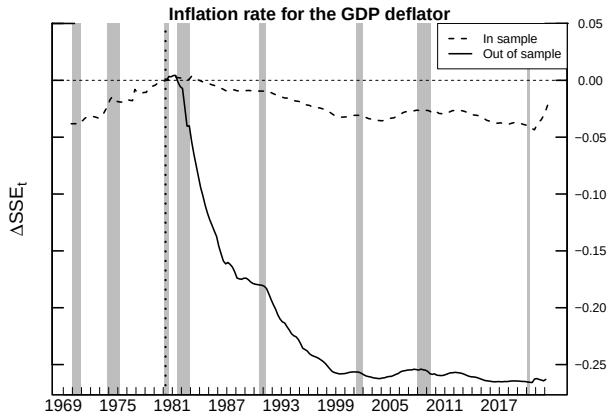
When we include an intercept in our regressions, the predictive performance of the models typically deteriorates except when there is a strong mean bias in the data to start with. As an illustration, consider the use of the Bordalo et al. (2020) model to predict individual inflation forecast errors using revisions at the individual level. Figure 4 contrasts the OOS performance of that model without an intercept (left panel) and with an intercept (right panel), for forecast errors of inflation based on the GDP deflator. Without an intercept, the model is able to outperform the null of rational expectations by a modest margin. But with an intercept, the performance is dismal. The reason is that the intercept is a regressor that has no cross-sectional variation and operates purely on the time-series dimension of the data, trying to fit mean bias in the average forecast. Because, as previously documented in Table 2, a mean bias model fits professional inflation forecasts poorly, this model inherits that bad performance. The online appendix documents that similar patterns hold for all models covered in this paper: Adding an intercept improves performance when the corresponding mean bias model performs well, while it deteriorates performance when the corresponding mean bias model performs poorly.

Figure 4: Effect of an intercept term on prediction (using revisions, individual level).

(a) Without intercept.



(b) With intercept.



Note: Dashed and solid lines show cumulative squared errors for the in-sample regression ΔSSE_t^{IS} and the out-of-sample regression ΔSSE_t^{OOS}, respectively, expressed as a fraction of the total sum of squared forecast errors over the evaluation period. Dotted vertical lines mark the end of the training period and the beginning of the evaluation period. An increase in a line indicates better performance of the behavioral model; a decrease in a line indicates better performance of the rational model. Shaded areas represent NBER recessions.

6.4 Data transformations

In our tests, we have transformed macroeconomic data following the conventions of the literature. In particular, we convert level forecasts of macroeconomic aggregates in the SPF into growth rate forecasts. Our results are robust to a range of data transformations, including taking log differences instead of working with growth rates in percentage points; working with quarter-over-quarter growth rates instead of annualized growth rates; and working with growth rates between the quarter of the forecast horizon and the previous quarter instead of growth rates between the quarter of the forecast horizon and the quarter before the survey date.[10]

7 Conclusion

This paper has shown that many models of biases in expectations documented in the literature are not robust to out-of-sample tests. These models seem unstable and would

[10] Detailed results of these additional tests are available upon request.

not have helped a forecaster to improve their predictions in real time. This general finding holds for professional forecasters and households.

However, there are some notable exceptions to this finding. First, interest rate forecasts display a stable mean bias that can be used to greatly improve forecasts out-of-sample. Second, there is some evidence for mean bias and autocorrelation in household inflation expectations. Third, individual expectations of professional forecasters display excess dispersion from the consensus for every variable and forecast horizon available. This excess dispersion is even more striking in inflation expectations in household surveys.

We hope that our findings will be useful to researchers in behavioral macroeconomics, where facts about deviations from rational expectations abound and models are validated by their success in matching moments corresponding to these facts. Our out-of-sample tests provide a simple and natural way to focus on those facts are most robust in the data and the associated research questions. These are: Why did forecasters systematically overpredict interest rates for decades? Why have household inflation expectations been so high for so long? And why do people disagree so much about the future, seemingly ignorant of the benefits of forecast combination?

References

Afrouzi, Hassan and Laura Veldkamp, "Biased Inflation Forecasts," 2019 Meeting Papers 894, Society for Economic Dynamics 2019.

Andolfatto, David, Scott Hendry, and Kevin Moran, "Are inflation expectations rational?," *Journal of Monetary Economics*, 2008, *55* (2), 406–422.

Angeletos, George-Marios, Fabrice Collard, and Harris Dellas, "Quantifying Confidence," *Econometrica*, 2018, *86* (5), 1689–1726.

_ , Zhen Huo, and Karthik A. Sastry, "Imperfect Macroeconomic Expectations: Evidence and Theory," *NBER Macroeconomics Annual*, 2021, *35*, 1–86.

Bhandari, Anmol, Jaroslav Borovička, and Paul Ho, "Survey Data and Subjective Beliefs in Business Cycle Models," Working paper 2022.

Bianchi, Francesco, Sydney C. Ludvigson, and Sai Ma, "Belief Distortions and Macroeconomic Fluctuations," *American Economic Review*, July 2022, *112* (7), 2269–2315.

Bonham, Carl S. and Douglas C. Dacy, "In Search of a "Strictly Rational" Forecast," *Review of Economics and Statistics*, 1991, *73* (2), 245–253.

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

Bürgi, Constantin and Julio Ortiz, "Overreaction Through Expectation Smoothing," Working paper 2022.

Clark, Todd E. and Kenneth D. West, "Approximately normal tests for equal predictive accuracy in nested models," *Journal of Econometrics*, 2007, *138* (1), 291–311.

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

Dovern, Jonas, Ulrich Fritsche, Prakash Loungani, and Natalia Tamirisa, "Information rigidities: Comparing average and individual forecasts for a large international panel," *International Journal of Forecasting*, 2015, *31* (1), 144–154.

Elliott, Graham, Ivana Komunjer, and Allan Timmermann, "Biases in Macroeconomic Forecasts: Irrationality or Asymmetric Loss?," *Journal of the European Economic Association*, 2008, *6* (1), 122–157.

Fama, Eugene F. and Kenneth R. French, "Dividend yields and expected stock returns," *Journal of Financial Economics*, 1988, *22* (1), 3–25.

Farmer, Leland, Emi Nakamura, and Jon Steinsson, "Learning About the Long Run," Working Paper 29495, National Bureau of Economic Research November 2021.

Gemmi, Luca and Rosen Valchev, "Biased Surveys," Working paper., Boston College 2021.

Goyal, Amit, Ivo Welch, and Athanasse Zafirov, "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction II," Working paper 2021.

Hajdini, Ina and Andre Kurmann, "Predictable Forecast Errors in Full-Information Rational Expectations Models with Regime Shifts," Working paper 2022.

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

Lettau, Martin and Stijn Van Nieuwerburgh, "Reconciling the Return Predictability Evidence: The Review of Financial Studies: Reconciling the Return Predictability Evidence ," *Review of Financial Studies*, 12 2007, *21* (4), 1607–1652.

Malmendier, Ulrike and Stefan Nagel, " Learning from Inflation Experiences," *Quarterly Journal of Economics*, 10 2015, *131* (1), 53–87.

McElroy, Tucker and Simon Sheng, "Augmented Information Rigidity Test," Working paper 2022.

Mincer, Jacob A and Victor Zarnowitz, "The evaluation of economic forecasts," in "Economic forecasts and expectations: Analysis of forecasting behavior and performance," NBER, 1969, pp. 3–46.

Nagel, Stefan and Zhengyang Xu, "Dynamics of Subjective Risk Premia," Working Paper 29803, NBER February 2022.

Nordhaus, William D., "Forecasting Efficiency: Concepts and Applications," *The Review of Economics and Statistics*, 1987, *69* (4), 667–674.

Pearce, Douglas K., "Short-term Inflation Expectations: Evidence from a Monthly Survey: Note," *Journal of Money, Credit and Banking*, 1987, *19* (3), 388–395.

Pfäuti, Oliver and Fabian Seyrich, "A Behavioral Heterogeneous Agent New Keynesian Model," Working paper 2022.

Rungcharoenkitkul, Phurichai and Fabian Winkler, "The Natural Rate of Interest Through a Hall of Mirrors," FEDS working paper 2022-010, Board of Governors of the Federal Reserve System 2022.

Sargent, Thomas J., "Rational Expectations," in David R. Henderson, ed., *The Concise Encyclopedia of Economics*, Liberty Fund, 2007.

Timmermann, Allan, "Forecast Combinations," in G. Elliott, C. Granger, and A. Timmermann, eds., *Handbook of Economic Forecasting*, Vol. 1, Elsevier, 2006, chapter 4, pp. 135–196.

Welch, Ivo and Amit Goyal, "A Comprehensive Look at The Empirical Performance of Equity Premium Prediction," *Review of Financial Studies*, 03 2007, *21* (4), 1455–1508.

Winkler, Fabian, "The role of learning for asset prices and business cycles," *Journal of Monetary Economics*, 2020, *114*, 42–58.

A Additional results

In this appendix, we report results of additional OOS tests using rolling window regressions, alternative forecast horizons, an added intercept, and a different way of transforming the raw data.

Table (A1) contains results from the consensus tests in Sections 4.1 and 4.2 but using rolling window regressions with a window length of 40 periods, i.e. ten years. The critical values for the bootstrap have been recomputed for the rolling window version of the ΔSSE_T statistic. Across all tests, the predictive OOS performance of the rolling regressions is worse than our benchmark recursive window regressions: The added noise from the shorter window length outweighs the benefits of accounting for time variation in the parameters. Table A2 contains the results for the individual models reported in Section 4.3 using rolling window regressions. Here, too, the predictive OOS performance of the rolling regressions is mostly worse than our benchmark recursive window regressions, although it is a little better for the Kohlhas-Walther model. Table A3 contains the results for household data reported in Section 5 using rolling window regressions. Here, the rolling window performance is almost identical to the recursive window performance.

The next two tables show results for different forecast horizons for a subsection of the variables in professional forecasts. Table A4 contains results at the consensus level while Table A5 contains results at the individual level. The results are broadly similar across forecast horizons. At times, the nowcast ($h = 0$) has a better performance than longer horizons and also crosses the significance thresholds of our bootstrap. Thus, nowcasts seem somewhat more predictable than longer-horizon forecasts. This may be due to higher power, as the forecast errors have lower variance at the short horizon.

Next, this appendix documents the OOS performance of some of the models when an intercept is added. Adding an intercept makes the OOS estimation noisier and lowers the power of the OOS test, but also allows to fit the data better if the true model contains an intercept. Table A6 shows results for professional forecasters at the consensus level and Table A7 shows results at the individual level. For the interest rate forecasts, adding an intercept to the models generally improves the OOS performance. This is consis-

tent with the good OOS fit of the mean bias model reported in the paper. For the other variables, the performance generally deteriorates, especially for the Coibion and Gorodnichenko (2015), Bordalo et al. (2020) and forecast combination models.

Table A1: Prediction of consensus professional forecast errors: rolling window regressions.

ΔSSE_T, OOS	(1) Revisions	(2) Mean bias	(3) Autocorrelation	(4) Mincer-Zarnovitz	(5) Nordhaus
Inflation (deflator)	-0.052	-0.254	0.039**	-0.557	0.051***
Inflation (CPI)	-0.028	-0.101	-0.089	-0.177	0.046***
Real GDP	-0.083	-0.067	-0.146	-0.190	0.083***
Industrial Production	-0.052	-0.021	-0.066	-0.076	-0.056
Nominal GDP	-0.108	-0.073	-0.484	-0.043**	0.028***
Unemployment rate	-0.046	-0.066	-0.156	-0.187	-0.150
Consumption	-0.019	-0.069	-0.075	0.014**	0.069***
Non-residential inv.	-0.015	-0.058	-0.034	-0.069	0.062***
Residential inv.	0.042***	-0.168	0.081**	-0.590	0.167***
Federal govt.	-0.063	-0.200	0.012	-0.019	-0.037
Non-federal govt.	-0.179	-0.047	-0.056	-0.090	0.048***
Housing starts	0.088***	-0.231	0.135***	-0.524	0.223***
Federal funds rate	0.193***	0.041**	-0.003	-0.080	0.200***
3-month yield	0.173***	0.109***	0.058**	-0.025**	0.197***
6-month yield	0.211***	0.153***	0.107***	-0.056	0.230***
1-year yield	0.193***	0.128**	0.041	-0.167	0.211***
2-year yield	0.143***	0.151***	0.010	-0.230	0.171***
10-year yield	0.034***	0.266***	-0.017	0.010**	0.058***
Aaa yield	0.024**	0.190***	-0.182	-0.136	0.082***
Baa yield	0.031**	0.385***	0.077**	0.185**	0.092***
1y-3m spread	-0.021	-0.124	-0.035	-0.094	-0.007
10y-2y spread	0.068***	-0.063	0.002	-0.135	0.072***
Aaa-Baa spread	-0.026	-0.094	0.059	-0.037	-0.129

Note: Each row shows cumulative squared errors ΔSSE_T in sample and out of sample. ***, ** and * represent rejection of the null hypothesis of no predictability of forecast errors at the 10, 5, and 1 percent level using bootstrapped critical values. Yield and spread variables are taken from BlueChip, other variables are taken from the SPF.

Table A2: Prediction of individual professional forecast errors: rolling window regressions.

ΔSSE_T, OOS	(1) BGMS	(2) Autocorrelation	(3) Mincer-Zarnovitz	(4) Nordhaus	(5) Kohlhas-Walther	(6) Forecast combination
Inflation (deflator)	0.001**	0.102***	-0.280	0.002***	-0.133	0.122***
Inflation (CPI)	-0.013	-0.054	-0.077	-0.023	-0.104	0.035***
Real GDP	0.034***	-0.166	-0.081	0.021***	0.054***	0.057***
Industrial Production	-0.024	-0.077	-0.044	-0.005	0.010**	0.049***
Nominal GDP	0.006***	-0.340	-0.142	0.001**	-0.012	0.057***
Unemployment rate	-0.023	-0.187	-0.135	0.007***	-0.050	0.016***
Consumption	0.077***	-0.112	-0.095	0.007***	0.021**	0.036***
Non-residential inv.	0.008**	-0.034	-0.042	-0.002	0.031**	0.059***
Residential inv.	-0.011	0.084***	-0.072	-0.015	0.009	0.099***
Federal govt.	0.081***	0.066***	-0.276	-0.042	-0.005	0.126***
Non-federal govt.	0.117***	0.025***	0.213***	0.005***	-0.058	0.169***
Housing starts	0.008***	0.212***	-0.560	-0.012	-0.220	0.100***
Federal funds rate	0.111***	0.040***	-0.065	0.077***	0.184***	0.033***
3-month yield	0.087***	0.097***	0.004	0.057***	0.252***	0.041***
6-month yield	0.105***	0.117***	-0.064	0.073***	0.267***	0.030***
1-year yield	0.084***	0.068***	-0.173	0.062***	0.204***	0.039***
2-year yield	0.060***	0.053***	-0.225	0.035***	0.226***	0.042***
10-year yield	0.007**	0.017**	-0.004	-0.013	0.257***	0.064***
Aaa yield	-0.004	-0.070	-0.145	-0.011	0.102***	0.063***
Baa yield	-0.008	0.142***	-0.052	-0.050	0.285***	0.115***
1y-3m spread	0.077***	-0.016	0.327***	0.100***	0.010	0.079***
10y-2y spread	-0.006	0.004	-0.062	-0.014	-0.034	0.053***
Aaa-Baa spread	0.007	-0.096	0.293***	-0.071	-0.006	0.066***

Note: Each row shows cumulative squared errors ΔSSE_T in sample and out of sample for a number of predictive models of forecast errors. ***, ** and * represent rejection of the null hypothesis of no predictability of forecast errors at the 10, 5, and 1 percent level using bootstrapped critical values.

Table A3: Prediction of household forecast errors for inflation: rolling regressions.

	(1)	(2)	(3)	(4)	(5)
ΔSSE_T, OOS	Mean bias	Revisions	Autocorrelation	Mincer-Zarnovitz	Forecast combination
Michigan mean	0.120***	-0.008	-0.165***	-0.139	–
Michigan median	-0.089	-0.018	-0.694	-0.358	–
Michigan ind.	0.007***	–	–	0.822***	0.889***
SCE mean	0.178	-0.013	-0.212	-0.364	–
SCE median	-0.280	0.048**	-0.140	-0.357	–
SCE ind.	-0.053	–	–	0.676***	0.803***

Note: Each row shows cumulative squared errors ΔSSE_T in sample and out of sample for a number of predictive models of forecast errors. ***, ** and * represent rejection of the null hypothesis of no predictability of forecast errors at the 10, 5, and 1 percent level using bootstrapped critical values. For the bootstrap on SCE data, we restrict the lag length of the MA process of u_t to 3. Forecast revisions are proxied by lagged forecasts.

Table A4: Prediction of consensus professional forecast errors: alternative horizons.

		(1)	(2)	(3)	(4)	(5)
	h	Revisions	Mean bias	Autocorrelation	Mincer-Zarnovitz	Nordhaus
Inflation (deflator)	0	0.008**	-0.035	0.040***	-0.039	-0.027
	1	-0.078	-0.140	0.078***	-0.128	0.047***
	2	-0.001	-0.270	0.156***	-0.282	0.066***
Real GDP	0	-0.225	-0.017	-0.028	-0.054	-0.079
	1	-0.195	-0.019	-0.003	-0.094	-0.105
	2	-0.160	-0.033	-0.013	-0.137	-0.074
Industrial Production	0	-0.074	-0.027	0.077***	0.064***	-0.106
	1	-0.135	-0.016	-0.039	-0.031	-0.068
	2	-0.040	-0.006	-0.035	-0.017	-0.007
Unemployment rate	0	-1.622	0.041***	0.110***	0.088***	-0.169
	1	-0.369	-0.003	-0.098	0.002	-0.180
	2	-0.367	-0.014	-0.057	-0.021	-0.126
Housing starts	0	0.090***	-0.016	0.051***	0.013**	0.071***
	1	0.018**	-0.028	-0.005	-0.039	0.147***
	2	0.066***	-0.037	0.059***	-0.066	0.205***
3-month yield	0	0.029**	0.439***	0.387***	0.508***	0.144***
	1	0.126***	0.160***	0.127***	0.180***	0.186***
	2	0.160***	0.121***	0.151***	0.112***	0.217***
10-year yield	0	0.087***	0.082***	0.004	0.037***	0.017**
	1	0.012	0.132***	-0.027	0.030**	0.027**
	2	-0.032	0.208***	-0.039	0.061**	0.042***
10y-2y spread	0	0.046***	-0.013	0.082***	-0.024	0.030**
	1	0.028**	-0.023	0.032**	-0.035	0.055***
	2	0.071***	-0.046	0.024	-0.050	0.09***

Note: Each row shows cumulative squared errors for the in-sample regression ΔSSE_T^{IS} and out-of-sample regression ΔSSE_T^{OOS}, respectively. All series are scaled by SSE_T^R, so that values correspond to the fraction of the mean squared forecast error predicted by the behavioral model. ***, ** and * represent rejection of the null hypothesis of no predictability of forecast errors at the 10, 5, and 1 percent level using bootstrapped critical values. Yield and spread variables are taken from BlueChip, other variables are taken from the SPF.

Table A5: Prediction of individual professional forecast errors: alternative horizons.

	h	(1) BGMS	(2) Autocorrelation	(3) Mincer-Zarnovitz	(4) Nordhaus	(5) Kohlhas-Walther	(6) Forecast combination
Inflation (deflator)	0	0.129***	0.080***	0.014**	0.006***	-0.015	0.041***
	1	0.038***	0.133***	-0.116	0.010***	-0.038	0.076***
	2	0.008***	0.161***	-0.279	0.007***	-0.091	0.117***
Real GDP	0	-0.069	-0.006	-0.063	-0.002	-0.016	0.018***
	1	0.038***	-0.024	0.062***	0.005***	0.006**	0.024***
	2	0.023***	-0.032	0.034**	0.004***	0.017**	0.048***
Industrial Production	0	0.013***	0.046***	0.015**	-0.015	-0.027	0.009***
	1	-0.022	-0.020	0.010	-0.008	-0.012	0.020***
	2	-0.012	-0.011	0.024**	-0.010	-0.007	0.044***
Unemployment rate	0	-0.413	0.058***	0.074***	-0.081	0.022***	0.010***
	1	-0.217	-0.111	0.022***	-0.083	0.000	-0.007
	2	-0.202	-0.084	0.012	-0.065	-0.010	0.003***
Housing starts	0	-0.001	0.087***	-0.015	-0.007	0.007**	0.038***
	1	-0.010	0.036***	-0.034	-0.006	-0.014	0.066***
	2	-0.002	0.148***	-0.057	-0.009	-0.018	0.095***
3-month yield	0	-0.008	0.217***	0.315***	0.053***	0.287***	0.025***
	1	0.045***	0.123***	0.162***	0.065***	0.181***	0.012***
	2	0.067***	0.153***	0.079***	0.059***	0.170***	0.027***
10-year yield	0	-0.001	0.020***	0.008***	-0.004	0.012***	0.005***
	1	-0.006	-0.019	0.028***	-0.008	0.062***	0.024***
	2	-0.010	-0.004	0.051***	-0.012	0.095***	0.042***
10y-2y spread	0	0.014***	0.020***	0.010***	-0.008	0.001	0.009***
	1	-0.011	0.040***	0.014***	-0.008	0.002	0.020***
	2	-0.004	0.037***	0.013***	-0.006	-0.005	0.037***

Note: Each row shows cumulative squared errors for the in-sample regression ΔSSE_T^{IS} and out-of-sample regression ΔSSE_T^{OOS}, respectively. All series are scaled by SSE_T^R, so that values correspond to the fraction of the mean squared forecast error predicted by the behavioral model. ***, ** and * represent rejection of the null hypothesis of no predictability of forecast errors at the 10, 5, and 1 percent level using bootstrapped critical values. Yield and spread variables are taken from BlueChip, other variables are taken from the SPF.

Table A6: Prediction of consensus professional forecast errors: adding an intercept.

ΔSSE_T, OOS	(1) Revisions	(2) Autocorrelation	(3) Nordhaus
Inflation (deflator)	-0.143	-0.084	0.001
Inflation (CPI)	-0.053	-0.106	0.005
Real GDP	-0.145	-0.021	-0.029
Industrial Production	0.032**	0.013	0.032***
Nominal GDP	-0.157	-0.070	-0.036
Unemployment rate	-0.274	-0.064	-0.124
Consumption	-0.049	-0.152	-0.076
Non-residential inv.	-0.125	-0.136	0.014**
Residential inv.	-0.088	-0.097	0.111***
Federal govt.	-0.075	-0.091	0.007
Non-federal govt.	-0.134	-0.076	0.089***
Housing starts	0.021**	0.075***	0.206***
Federal funds rate	0.211***	0.108**	0.213***
3-month yield	0.238***	0.174***	0.212***
6-month yield	0.266***	0.198***	0.251***
1-year yield	0.243***	0.149***	0.238***
2-year yield	0.192***	0.139**	0.191***
10-year yield	0.283***	0.27***	0.120***
Aaa yield	0.226***	0.205***	0.127***
Baa yield	0.382***	0.411***	0.162***
1y-3m spread	-0.144	-0.332	0.030**
10y-2y spread	0.017	-0.070	0.079***
Aaa-Baa spread	-0.199	-0.105	-0.180

Note: Each row shows cumulative squared errors ΔSSE_T in sample and out of sample. ***, ** and * represent rejection of the null hypothesis of no predictability of forecast errors at the 10, 5, and 1 percent level using bootstrapped critical values. Yield and spread variables are taken from BlueChip, other variables are taken from the SPF.

Table A7: Prediction of individual professional forecast errors: adding an intercept.

ΔSSE_T, OOS	(1) BGMS	(2) Autocorrelation	(3) Nordhaus	(4) Forecast combination
Inflation (deflator)	-0.263	-0.074	-0.062	-0.157
Inflation (CPI)	-0.085	-0.088	-0.044	-0.038
Real GDP	0.013	-0.035	0.024***	0.042**
Industrial Production	0.013	-0.001	0.014***	0.06***
Nominal GDP	-0.020	-0.084	0.007**	0.037**
Unemployment rate	-0.136	-0.075	-0.066	-0.012
Consumption	0.027	-0.189	0.004	0.009
Non-residential inv.	-0.091	-0.155	-0.011	-0.008
Residential inv.	-0.084	-0.038	0.003	0.033
Federal govt.	0.029**	0.019**	-0.014	0.082***
Non-federal govt.	0.091***	0.011	-0.004	0.139***
Housing starts	-0.029	0.18***	-0.004	0.065**
Federal funds rate	0.111***	0.111***	0.062***	0.117***
3-month yield	0.179***	0.188***	0.055***	0.209***
6-month yield	0.197***	0.175***	0.104***	0.182***
1-year yield	0.163***	0.140***	0.093***	0.165***
2-year yield	0.162***	0.139***	0.022***	0.205***
10-year yield	0.225***	0.218***	0.012***	0.293***
Aaa yield	0.104***	0.099***	0.036***	0.167***
Baa yield	0.291***	0.308***	0.039***	0.396***
1y-3m spread	-0.003	-0.162	0.103***	0.002
10y-2y spread	-0.051	-0.051	-0.028	-0.001
Aaa-Baa spread	-0.113	-0.531	-0.159	-0.034

Note: Each row shows cumulative squared errors ΔSSE_T in sample and out of sample for a number of predictive models of forecast errors. ***, ** and * represent rejection of the null hypothesis of no predictability of forecast errors at the 10, 5, and 1 percent level using bootstrapped critical values.