Can You Improve Upon the GDP Forecasts of Professional Forecasters Using Information About Monetary Policy?

Dean Croushore[1]

Abstract

In this paper, I examine the forecast errors of macroeconomic forecasters to see whether or not their forecasts are efficiently using information about monetary policy. The goal is to investigate, using real-time data, previous research that has found inefficiency in forecasts with respect to monetary policy. I use a real-time data set to investigate the relationship between GDP forecast errors and changes in monetary policy both in-sample and with out-of-sample methods. Out-of-sample results show that exploiting inefficiency is difficult in real time.

Keywords: evaluating forecasts, macroeconomic forecasting, real-time data, forecast efficiency, forecast improvement

[1]dcrousho@richmond.edu, Economics Department, Robins School of Business, 102 UR Drive, University of Richmond, VA, 23173, USA

1. Introduction

Do forecasters optimally change their forecasts of GDP growth in response to changes in monetary policy? If so, then forecasters are efficient in their analysis of the effects of monetary policy on GDP growth. If not, can a researcher help forecasters make better forecasts by incorporating information about monetary policy in a superior manner?

Tests of the efficiency of GDP forecasts with respect to monetary policy have been conducted by a few papers in the literature but mostly using in-sample methods and based on final, revised data. In this paper, I examine the question in a more convincing manner, using real-time data to account more accurately for data revisions, using out-of-sample methods to examine the robustness of in-sample results, and exploring how inefficiency changes over time. The key question is: could a researcher improve on GDP forecasts in real time using information about monetary policy?

There is a vast literature on the evaluation of forecasts. Point forecasts are evaluated most often using tests of unbiasedness and efficiency. The literature in this area was summed up most clearly by Clark and Mertens (2024), who suggest that forecasts from surveys of professional forecasters are "competitive (albeit not fully optimal) predictors of future outcomes." Recent research has suggested a number of problems with the forecasts of professionals. Theoretical reasons for forecast inefficiency include noisy-information models in which agents exhibit rational inattention because information-processing constraints lead to forecast inefficiency, as in Sims (2003), and sticky-information models, in which there are informational rigidities, as proposed by Mankiw and Reis (2002), where information disperses slowly to agents. Those theories were and tested and contrasted by Coibion and Gorodnichenko (2012), who showed evidence that was more supportive of noisy-information models, and certainly was inconsistent with rational-expectations models. Bordalo et al. (2020) found that the consensus of forecasts from a survey under-react to news, while individual forecasters over-react, and developed a model of dispersed information to ex-

plain it. Clements (2022) showed that individual forecasters are inefficient in their use of information. Bianchi et al. (2022) found that individual forecasters suffer from belief distortions but that artificial intelligence algorithms can be used to improve their forecasts. Eva and Winkler (2023), however, found that the research on forecast errors is not very robust and cannot be used to improve on the forecasts in a true real-time out-of-sample experiment. I follow the recent structure of Croushore (2025) to explore whether or not the forecasts can be improved out-of-sample in real-time using alternative methods of dealing with data revisions (see Koenig et al. (2003)), accounting for structural instability (see Rossi and Sekhposyan (2010).

The literature suggests that GDP forecasts may not respond appropriately to shocks to monetary policy. Several papers, Ball and Croushore (2003) and Rudebusch and Williams (2009), showed that forecasters do not modify their GDP forecasts properly when monetary policy changes. Ball and Croushore (2003) found that observable changes in monetary policy caused forecasters to change their forecasts of GDP growth, but not enough. That is, tighter policy caused forecasters to revise down their GDP forecasts, but GDP growth in fact declined even more than the forecasters thought.[2] They found that the change over the last year in the real fed funds rate was correlated with one-year-ahead GDP forecast errors, in-sample. Since the measure of monetary policy was known to forecasters when their forecasts were made, the results imply that the forecasts are inefficient. Rudebusch and Williams (2009) used a different measure of monetary policy, the yield spread, and found a similar result: quarterly real GDP forecast errors were correlated with the lagged yield spread, which was also in the information set of the forecasters when their forecasts were made.

In this paper, I explore the robustness of the Ball and Croushore (2003) and

[2]However, Ball and Croushore (2003) found that the forecasts of inflation were efficient with respect to monetary policy, so I do not investigate inflation forecasts in this paper.

Rudebusch and Williams (2009) results when the analysis includes real-time out-of-sample tests. I examine whether a researcher could have used data available to the forecasters to make better GDP forecasts using the available data on monetary policy. The result is that forecast improvement is difficult, despite in-sample evidence of inefficiency.

2. Data

In this paper, I examine forecasts from the Survey of Professional Forecasters (SPF), which is widely studied.[3] I examine forecasts for real output growth, measured as GNP before 1992 and GDP from 1992 on. The forecasts are made quarterly and the survey asks the respondents to forecast the growth of real output in the current quarter and each of the following four quarters. I examine each of the quarterly annualized forecasts as well as the average output growth forecast over the next four quarters.

Quarterly forecasts for output growth (at an annualized rate) are calculated as in Equation (1):

$$y^f_{t,t+h} = (((\frac{Y^f_{t,t+h}}{Y^f_{t,t+h-1}})^4) - 1) \times 100\%, \tag{1}$$

where $h = 0$, 1, 2, 3, and 4, and $Y^f_{t,t+h}$ is the level of the output forecast made at date t for date $t+h$, using data on output through date $t-1$.

For testing purposes, I compare those forecasts to realized values, which are calculated as

$$y_{t+h} = ((\frac{Y_{t+h}}{Y_{t+h-1}})^4 - 1) \times 100\%. \tag{2}$$

The forecast error is the realized value of the growth rate minus the forecast

$$e_{t,t+h} = y_{t+h} - y^f_{t,t+h}. \tag{3}$$

[3]The SPF is the only quarterly survey of U.S. macroeconomic forecasters available at no charge, and has been produced on a quarterly basis since 1968. See Croushore and Stark (2019) for a historical discussion of the SPF and the research that uses it. Because of irregularities in the early years of the survey, I start the analysis from the first quarter survey in 1971.

In addition to quarterly forecasts, the SPF can also be used for annual forecasts, both from the current quarter to four-quarters ahead, and from the quarter prior to the forecast date to three-quarters ahead. The average annual output growth rate forecast over quarters t to $t+4$ is calculated in Equation (4):

$$y^{f4}_{t,t+4} = (\frac{Y^{f}_{t,t+4}}{Y^{f}_{t,t}} - 1) \times 100\%. \tag{4}$$

Realized values over the same period are

$$y^{4}_{t+4} = (\frac{Y_{t+4}}{Y_{t}} - 1) \times 100\%. \tag{5}$$

Thus forecast errors for average annual forecasts are equal to

$$e^{4}_{t} = y^{4}_{t+4} - y^{f4}_{t,t+4}. \tag{6}$$

Similarly, I can calculate the average annual forecast growth rate from quarters $t-1$ to $t+3$ by lagging Equation (4) by one quarter; similarly for the realized values and forecast errors.

A key question in the forecasting literature is which vintage of the data to use as the realized value in Equations (2) and (5). There are many alternatives and I explore differences across them, comparing initial realized values (the release at the end of the first month of the following quarter), to first-final realized values (the release at the end of the third month of the following quarter), to first-annual realized values (the release at the end of July of the following year in most years), to pre-benchmark realized values (the last release before a benchmark revision of the National Income and Product Accounts), to latest-available realized values (from the latest available vintage of data available when this research started, which was August 2024). If data revisions are small and unimportant, then the latest-available realized values are the best choice. However, the real-time literature, as summarized in Croushore (2011), shows that measures other than the latest-available realized value may be superior. The initial or first-final realized values have the advantage of being released not long

after a quarter ends, but the disadvantage of being based on very incomplete source data. Most analyses seem to be improved by using either the first-annual realized values, with fairly complete sourse data, or pre-benchmark realized values, which are the last vintage available under a consistent methodology prior to a benchmark revision.

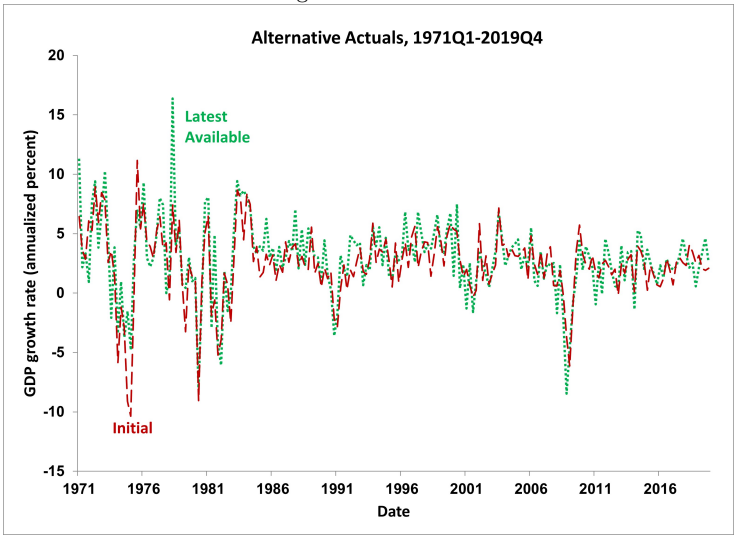
I obtain the alternative realized values from the Real-Time Data Set for Macroeconomists (RTDSM), which was created by Croushore and Stark (2001) and made available on the website of the Federal Reserve Bank of Philadelphia. The RTDSM provides data on real output and other major macroeconomic variables, as someone standing at the middle of any month from November 1965 to today would have viewed the data. The RTDSM lines up perfectly with the SPF in terms of data availability.

Figure 1 plots GDP growth rates for two of the alternative realized values, initial and latest available, from 1971Q1 to 2019Q4. You can see that the two series generally move together, but there are quarters when they differ substantially, in one case by over ten percentage points. Thus, forecast evaluation conclusions potentially differ significantly depending on the choice of realized values.

To visualize what the realized values and forecasts look like, Figure 2 shows a plot of the forecast for average annual output growth over the next four quarters (one-to-four-quarters ahead) and the initial realized value of GDP growth over the same horizon. Note that the graph ends prior to the COVID period, to avoid distortions caused by the large swings to GDP growth in 2020; so it uses forecasts from 1971Q1 to 2018Q4, with the corresponding realized values (ending in 2019Q4). As expected, the forecasts are a much smoother series than the object being forecast, and there are some large unanticipated shocks to output.

To provide a sense of the size of forecast errors, Figure 3 shows representative forecast errors based on the initial concept of realized values at quarterly horizons 0 and 4. The forecast errors are large and volatile, and they change signs

Figure 1: Alternative Realized Values



Note: The figure shows the quarterly realized values of GDP growth rates as calculated using Equation (2) based on two alternative concepts: initial and latest available. The graph ends prior to the COVID period, to avoid distortions caused by the large swings to GDP growth in 2020.

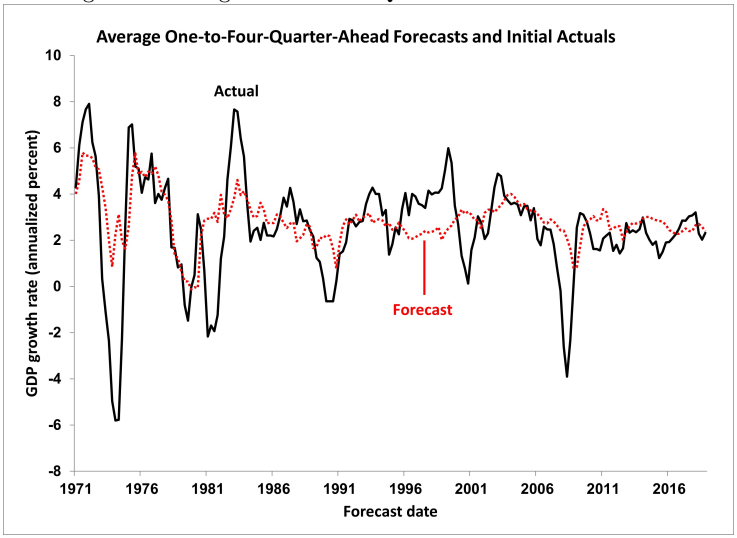
frequently, making them difficult to predict.

To examine whether measures of monetary policy might be used to improve GDP forecasts, I consider two alternative measures of monetary policy: the yield spread and changes in the real federal funds rate.[4] For the yield spread, I use the measure of Rudebusch and Williams (2009), which is the interest rate on 10-year Treasury notes minus the interest rate on 3-month Treasury bills, using the constant-maturity series for each security. For the change in the real federal funds rate, I use the Ball and Croushore (2003) measure, which is the change in the expected real federal funds rate over the past year.[5] Note that both the yield spread and the change in the real fed funds rate are available to

[4]Other measures of monetary policy, such as measures of monetary policy surprises from a VAR, are not available in real time, as they were not obviously in the information set of forecasters when they made their forecasts.

[5]Ball and Croushore (2003) examined alternatives to this measure and found that the results were not sensitive to the proxy used.

Figure 2: Average One-to-Four Quarter Ahead Forecasts and Initial Realizations



Note: The figure shows the forecast for average annual growth over the next four quarters and the initial realized value of GDP growth over the same horizon. The graph uses forecasts from 1971Q1 to 2018Q4, with the corresponding realized values (ending in 2019Q4).

the SPF forecasters at the time they make their forecasts. I am careful to use only data available to the forecasters in these efficiency tests.

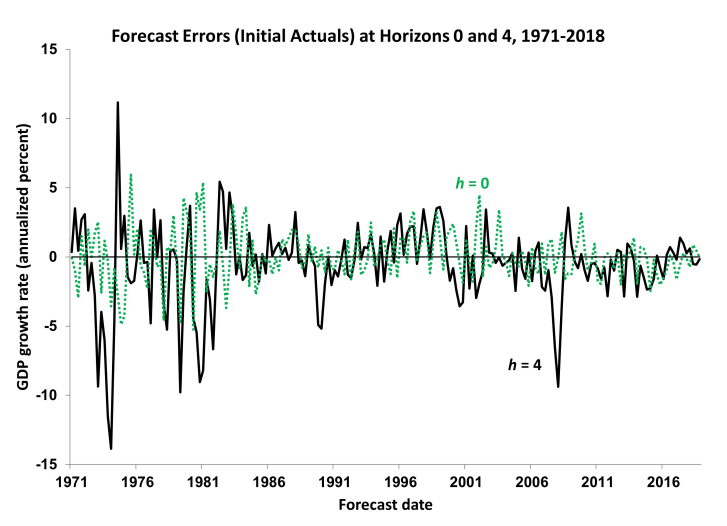
The two measures of monetary policy differ somewhat over time but their major movements are correlated, with the real fed funds rate measure having an inverse correlation with the spread measure, as you can see in Figure 4.

3. In-Sample Results

In this section, I investigate whether the two measures of monetary policy are correlated with forecast errors in-sample. The sample uses all SPF forecasts made from 1971Q1 to 2018Q4, so that four-quarter-ahead forecasts end before COVID begins in 2020.[6] In the analysis, in addition to extending the Rudebusch-Williams and Ball-Croushore results, I use each of their measures of

[6]Including the COVID period in the sample changes the results because of the huge swings in GDP growth in 2020 and 2021.

Figure 3: Forecast Errors at Horizons 0 and 4



Note: The figure shows the quarterly forecast errors for GDP growth rates as calculated using Equation (3) for two horizons: current quarter ($h = 0$) and four quarters ahead ($h = 4$), and using the initial data release as the realized value. The graph uses forecasts from 1971Q1 to 2018Q4, with the corresponding realized values (ending in 2019Q4 at the latest).

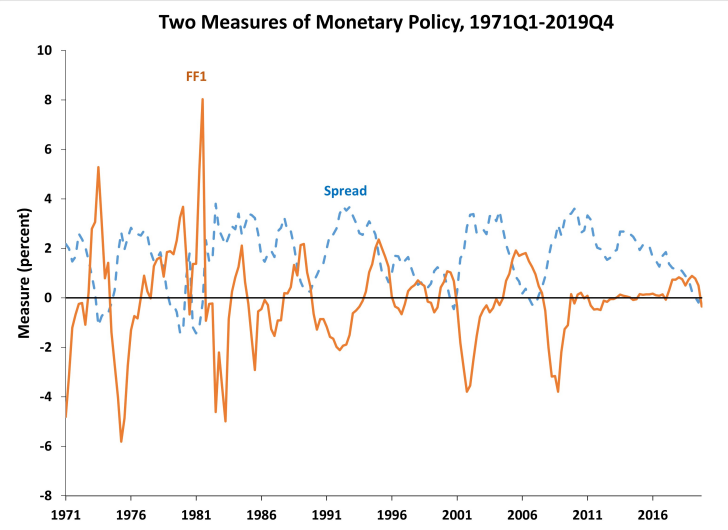
monetary policy on the other one's method, to see how robust they are.

Ball and Croushore (2003) looked at the forecast of the average growth rate of real output over the coming year, which I defined above as Equation (4), and compared it to the realized value, given by Equation (5). The forecast error is given by Equation (6).

In Table 1, I show the original published result from Ball and Croushore (2003), which was based on forecasts from 1968Q4 to 1995Q2, and updated results based on a longer sample period using forecasts from 1971Q1 to 2018Q4.[7] As the table shows, the Ball-Croushore results hold up well with an additional 23 years of data. However, using the Rudebusch-Williams term spread as the measure of

[7] Ball and Croushore (2003) started in 1968Q4 because they were not aware of the problems in the SPF data in the late 1960s. The other difference is that they used the median across forecasters but I use the mean, although the differences between mean and median forecasts in the SPF is trivial.

Figure 4: Two Measures of Monetary Policy



Note: The figure shows the two alternative measures of monetary policy that I use: the term spread between 10-year T-notes and 3-month T-bills (Spread) and the change in the real fed funds rate over the previous year ($FF1$).

monetary policy leads to insignificant results, even though the two variables, $FF1$ and S are highly correlated.

Table 1: Ball–Croushore Results and Update

Regression: $e_t^4 = \beta MP_{t-1} + \epsilon_t^4$

	Original FF1	Update FF1	Update S
MP	**−0.464**	**−0.365**	0.029
	(0.143)	(0.092)	(0.046)
χ^2 sig.	<0.01	<0.01	0.523
$\overline{R}^2$	0.20	0.10	-0.022

Notes: The table shows the original results, in the column headed "Original FF1" reported by Ball and Croushore (2003), the updated version headed "Update FF1", and using the spread, headed "Update S". Numbers in parentheses are HAC standard errors to adjust for overlapping observations. Bold numbers indicate coefficients that are statistically significantly different from zero at the 5 percent level. The original sample period covers SPF forecasts made from 1968Q4 to 1995Q2, whereas the updated results use SPF forecasts from 1971Q1 to 2018Q4.

Following a similar procedure, we run the Rudebusch-Williams in-sample regressions, as shown in Table 2.

The results shown in Table 2 are broadly consistent across sample periods and measures of monetary policy and confirm the in-sample results that observable changes in monetary policy are significantly related with real output forecast errors.[8]

Most of the time, the other coefficients included in these regressions, the constant term and the lagged forecast term, are not statistically significant in the in-sample regressions. Because our ultimate goal is to use the regressions to make better forecasts, the parsimony principle suggests removing those terms from the regressions and using a simpler structure. So, I run a regression of each of the forecast errors for the seven horizons and four different measures of realizations for each of the two different measures of monetary policy. The regression is simply:

$$e_{t,t+h} = \beta MP_{t-1} + \epsilon_t^h, \tag{7}$$

where MP_{t-1} is one of the measures of monetary policy through date $t-1$ (known to forecasters making their forecasts at date t) and $e_{t,t+h}$ is a forecast error from Equation (3) or (6). The results are summarized in Table 3.

In Table 3, we see that about half of all the cases (denoted "S") show a statistically significant coefficient (p-value ≤ 0.05) in regression Equation (7), which suggests that forecasters are not using information about monetary policy efficiently in forming their forecasts. The coefficients on monetary policy are most often significant at longer horizons, which is consistent with the literature allowing for a lag in the effect of monetary policy on output. In terms of the alternative measures of realized values, the coefficients on monetary policy are more often significant for using first annual or pre-benchmark realized values.

[8]The difference in the signs on the yield spread coefficient arose from an innocuous sign flip in the Rudebusch and Williams (2009) paper.

Table 2: Rudebusch–Williams Results and Replication

Regression: $e_{t+h|t-1} = \alpha + \beta y^e_{t+h|t-1} + \gamma S_{t-1} + \epsilon_{t+h|t-1}$

	Original S	Update S	Update $FF1$
	Current-quarter forecast		
Constant	−0.04	0.21	0.31
SPF forecast	0.08	−0.04	−0.02
Yield spread	−0.10	0.09	0.11
F-test (p-value)	0.34	0.27	0.27
	One-quarter-ahead forecast		
Constant	−0.52	−0.12	0.36
SPF forecast	−0.19	−0.23	−0.14
Yield spread	**−0.65**	**0.42**	**−0.27**
F-test (p-value)	0.01	0.14	0.20
	Two-quarter-ahead forecast		
Constant	−0.14	−0.14	0.37
SPF forecast	**−0.50**	−0.36	−0.29
Yield spread	**−0.88**	**0.55**	**−0.36**
F-test (p-value)	0.00	0.10	0.03
	Three-quarter-ahead forecast		
Constant	−0.70	−0.33	0.73
SPF forecast	−0.31	−0.36	−0.29
Yield spread	**−0.76**	**0.55**	**−0.47**
F-test (p-value)	0.02	0.15	0.00
	Four-quarter-ahead forecast		
Constant	−0.33	−1.42	0.44
SPF forecast	−0.37	−0.12	−0.33
Yield spread	**−0.68**	**0.72**	**−0.53**
F-test (p-value)	0.00	0.02	0.00

Notes: The table shows the original results reported by Rudebusch and Williams (2009) in the column headed "Original S", the updated version headed "Update S", and using the lagged change in the real Fed funds rate, headed "Update $FF1$". Bold numbers indicate coefficients that are statistically significantly different from zero at the 5 percent level. HAC standard errors are used to account for overlapping observations but are not shown to conserve space. The original sample period covers SPF forecasts made from 1968Q4 to 2007Q1, whereas the updated results use SPF forecasts from 1971Q1 to 2018Q4.

Table 3: In-Sample Results for Monetary Policy

$e_{t,t+h} = \beta MP_{t-1} + \epsilon_t^h$

Horizon	0	1	2	3	4	1-4	0-3
Realized Value							
initial	x x	x M	x S	x S	x S	x S	x S
first final	M x	x M	x S	x S	x S	x S	x S
first annual	M x	x S	x S	M S	x S	x S	x S
pre-benchmark	M x	x S	x S	M S	x S	x S	x S
latest-available	S x	M x	x x	S x	M S	S S	S S

Note: Results of test of null hypothesis that $\beta = 0$: x means p-value > 0.10; M means $0.05 < p$-value < 0.10; S means p-value ≤ 0.05. First term: yield spread; second term: lagged change in real fed funds rate. The sample uses SPF forecasts from 1971Q1 to 2018Q4. Standard errors are adjusted following the Newey and West (1987) procedure.

4. Forecast Improvement

The in-sample results are based on the full sample of forecasts made from 1971Q1 to 2018Q4. They do not show how a researcher standing at different points in time would have perceived the inefficiency regressions. Croushore (2025) suggests three approaches for viewing inefficiency in real time: Continuously-Updated, Benchmark-Consistent, and Vintage-Specific. In this paper, I only consider the Vintage-Specific approach; results of the other approaches are discussed in the Appendix. I look at the forecast-rationality statistics of Rossi and Sekhposyan (2016), using 5-year and 10-year rolling windows, with the idea being that even if the full-sample in-sample results do not show inefficiency, that may be because the inefficiencies in short periods offset each other. The method helps identify the periods of inefficiency.

In-sample Results for Vintage-Specific Approach. In the Vintage-Specific approach, we think about a researcher assuming that forecasters at each date look at data vintages with similar ages, with the idea that the data-generating

process (DGP) differs across concepts of realized values. In particular, especially for forecasting initial values, using just the initial release values in the forecasting model might be appropriate. The assumption is that the researcher and forecasters view the DGP as relating initial releases to each other over time.

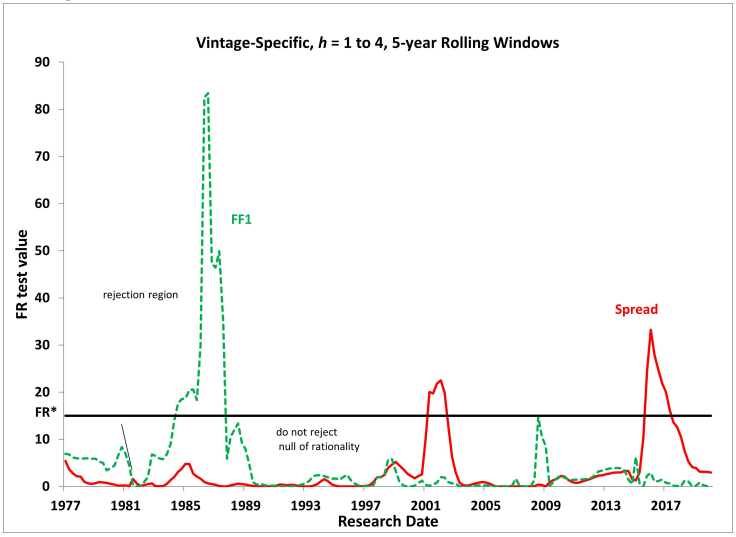
So, imagine a researcher standing in 1976Q1, evaluating the current-quarter forecasts from the SPF from 1971Q1 to 1975Q4 (a five-year window), using the initial-release realized values to analyze the forecasts. Then roll the exercise forward quarter by quarter, maintaining a five-year window each time. Do the same exercise for each of the two different measures of monetary policy and the seven different forecast horizons, allowing for a longer lag in data availability as the horizon lengthens. For each five-year window, calculate the forecast-rationality statistic and compare it with the critical value from Rossi and Sekhposyan (2016). The forecast-rationality statistic is calculated at each research date, and I reject the null hypothesis of forecast rationality if any of the values exceeds the critical value for any research date.

I show the results here for the $h = 1$ to 4 horizon, with other results in the Appendix. Figure 5 shows that for this longer horizon with the Vintage-Specific approach using 5-year rolling windows, there are a number of rejections of forecast rationality. Repeating this exercise for 10-year rolling windows leads to rejections of forecast rationality only for the $FF1$ measure of monetary policy, as Figure 6 shows.

5. Forecast-Improvement Exercises for Inefficiency in Real Time

Given the in-sample results, I proceed to investigate the possibility of using the regression results from Equation (7) to improve upon the SPF forecasts in a simulated real-time out-of-sample exercise; I call this a forecast-improvement exercise (FIE). Taking the estimated $\hat{\beta}$, and recalling from Equation (3) that $e_{t,t+h} = y_{t+h} - y^f_{t,t+h}$, I create, at each date t, an improved forecast $y^i_{t,t+h}$, where

Figure 5: Forecast-Rationality Test Results, Vintage-Specific Approach, h = 1 to 4, 5-Year Rolling Windows



Notes: The figure shows the forecast-rationality test results using the Vintage-Specific approach with a 5-year rolling window and a horizon of 1 to 4 quarters. The forecast-rationality critical value is labeled FR*, and we reject forecast rationality if any value across the sample exceeds that threshold. Each line is labeled with the type of monetary policy used in Equation (7), where S is the yield spread and $FF1$ is the lagged change in real fed funds rate.

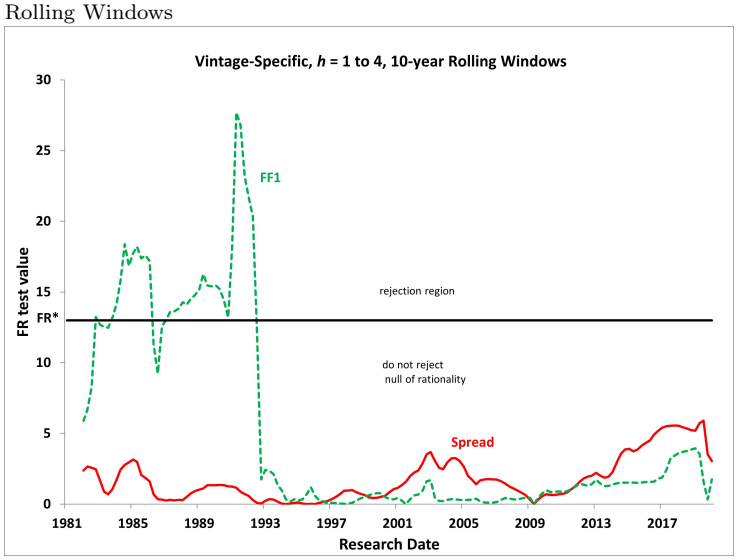
$$y^i_{t,t+h} = y^f_{t,t+h} + (\delta_t \times \hat{\beta} MP_{t-1}), \tag{8}$$

where the δ_t term is described below. The baseline case has δ_t equal to 1 for all t.

Using Equations (7) and (8), I simulate the activity of a real-time researcher, forming improved forecasts at each date based only on the real-time data and past forecast errors available at each date.[9] I calculate root-mean-squared-forecast errors ($RMSFE$s) for each different horizon and each different measure of monetary policy. I compare those $RMSFE$s to those of the SPF forecast,

[9]In the out-of-sample exercise, I use only data that the forecasters would have known in real time when the SPF survey results are released, using the data only up to the quarter prior to the SPF forecast. The coefficients of the regression are re-estimated at each date.

Figure 6: Forecast-Rationality Test Results, Vintage-Specific Approach, h = 1 to 4, 10-Year Rolling Windows



Notes: The figure shows the forecast-rationality test results using the Vintage-Specific approach with a 10-year rolling window and a horizon of 1 to 4 quarters.

dividing the $RMSFE$ of the attempt to improve on the survey by the $RMSFE$ of the SPF, to generate a relative root-mean-squared forecast error ($RRMSFE$). An $RRMSFE$ greater than one means the attempt to improve on the SPF forecasts actually made them worse, while an $RRMSFE$ less than one means the attempt to improve on the SPF succeeded.[10]

I consider four different versions of Equation (8). The baseline case has $\delta_t = 1$ for all t. This method does not account for estimation error in the coefficients, however. To account for estimation error, I could shrink the estimated coefficients towards zero, as suggested in the literature on forecast combination.[11] As a simple first pass, I shrink the coefficients by half, so that $\delta_t = 0.5$ for all t. (I leave it for future research to determine optimal shrinkage in these

[10] Statistical significance of differences between the surveys is tested using the Harvey et al. (1997) modified Diebold and Mariano (1995) test statistic of the corresponding null hypothesis.
[11] See Diebold and Pauly (1990) and Diebold and Lopez (1996).

forecast-improvement exercises.) An alternative is to use the information from the forecast-rationality tests to allow the δ_t term to vary over time. One possibility is to set the δ_t term to zero, if the FR-test value is less than the critical value at that date, or equal to one, if the FR-test value is greater than the critical value at that date. Another possibility is to use shrinkage with that method. So, I try all four adjustment methods to see how the results vary.

There are many results of these forecast-improvement exercises: 2 rolling window sizes (5-year and 10-year), 2 measures of monetary policy (S, $FF1$), 7 horizons (0, 1, 2, 3, 4, 1 to 4, 0 to 3), and 4 adjustment methods (full, full with shrinkage, FR-test based, FR-test based with shrinkage), for a total of 112 sets of results.

In what follows, I first focus on the 1-to-4 quarter-ahead horizon, with other results shown later. Table 4 shows the results for the Vintage-Specific approach for the 1-to-4 quarter horizon. The table shows the results in columns for the two different measures of monetary policy and the two differing rolling window lengths. The rows show the $RRMSFE$ for each different adjustment method to improve on the forecasts. The values in the table are the $RRMSFE$ for that case, with the p-values of the Diebold-Mariano test shown in square brackets below each $RRMSFE$. In 4 of the 16 cases the $RRMSFE$ is equal to exactly 1.000 because the FR test did not reject the null of efficiency, so there is no adjustment to the forecasts at all. In one case, using the yield spread with shrinkage in 10-year rolling windows, the $RRMSFE$ is slightly below one, though not statistically significant. In the remaining cases, 11 out of 16, the attempt to improve the forecasts actually makes them worse, although none of them are statistically significantly worse (at the 5 percent level). Note that the experiments shown in this table for the $FF1$ measure show the attempt to improve on the forecasts based on the Ball-Croushore results, but in 2 of the 8 cases the $RRMSFE$ equals one and in the other 6 cases the $RRMSFE$ exceeds one, though it is never statistically significantly above one.

Looking more generally across all the horizons, and not just at the 1-to-4 quarter

Table 4: $RRMSFEs$ and P-values for Forecast Improvement Exercises, Vintage-Specific approach with realized values = initial, $h = 1 - 4$

Monetary Policy Measure	S		$FF1$	
Window Size: Adjustment Method	5-yr	10-yr	5-yr	10-yr
All	1.103 [0.17]	1.015 [0.63]	1.064 [0.26]	1.025 [0.54]
All, shrink	1.023 [0.35]	0.999 [0.93]	1.012 [0.61]	1.000 [0.99]
$FR > cv$	1.063 [0.31]	1.000 [0.99]	1.003 [0.67]	1.019 [0.27]
$FR > cv$, shrink	1.020 [0.32]	1.000 [0.99]	1.000 [0.99]	1.005 [0.48]

Notes: The table shows relative-root-mean-squared errors ($RRMSFE$) and p-values of the Diebold-Mariano test [in square brackets] for forecasts in forecast-improvement exercises, using the Vintage-Specific approach with realized values = initial. The sample consists of one-year-ahead SPF forecasts made from 1971Q1 to 2018Q4.

horizon, Table 5 shows the results of all 112 permutations of 2 rolling window sizes, 2 measures of monetary policy, 7 horizons, and 4 adjustment methods. The table shows the number of cases in each of seven different ranges for $RRMSFE$ and the number of cases with a p-value of less than or equal to 0.05.

Table 5: Forecast Improvement Exercises Counts of Ranges, All Permutations, Vintage-Specific Approach

$RRMSFE$ range	$RRMSFE$s in range	p-value ≤ 0.05
1.10 to ∞	2	0
1.02 to 1.10	23	4
1.00 to 1.02	46	0
1.00 exactly	28	0
0.98 to 1.00	13	0
0.90 to 0.98	0	0
0.00 to 0.90	0	0

Notes: The table shows the numbers of cases in which the relative-root-mean-squared error ($RRMSFE$) falls within the given range, and the number of cases with p-values of the Diebold-Mariano test that are less than or equal to 0.05, across all 112 permutations of 2 rolling window sizes, 2 measures of monetary policy, 7 horizons, and 4 adjustment methods. The sample consists of SPF forecasts made from 1971Q1 to 2018Q4.

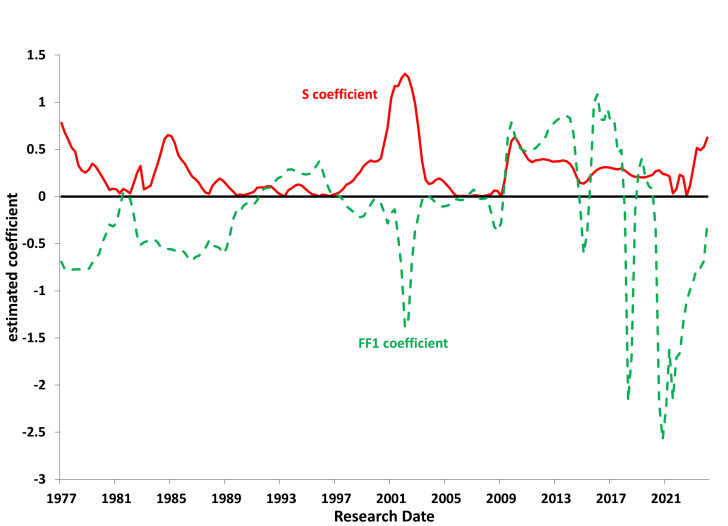
As Table 5 shows, there are no statistically significant improvements to the forecasts. In no cases out of 112 is there more than 2 percent improvement in the $RMSFE$. In just 13 of the 112 cases, there is a small improvement on the SPF forecasts of less than 2 percent of the SPF $RMSFE$, and the improvement is never statistically significant. In 28 cases out of 112, the FR test does not reject the null of efficiency, so there is no adjustment in the forecasts, and the $RRMSFE$ is exactly one. In the other 71 cases out of 112, which is 63 percent of all cases, the attempt to improve the forecasts makes them worse; they are statistically significantly worse in about 3.6 percent of all the cases. Of course, we might expect such results in about 5 percent of all cases by the random

nature of such experiments.[12]

6. Why Do Forecast-Improvement Exercises Fail?

Our analysis shows that improving upon the forecasts is possible, but never statistically significant, and improvement is rare. One reason is that, as Figures 5 and 6 show, there are not many violations of efficiency in-sample. In addition, the estimated coefficient on the measure of monetary policy changes over time, as Figure 7 shows. The sharp movements in the estimated coefficients on the measure of monetary policy may introduce noise into the improved forecasts, causing them to be worse than the original SPF forecasts.

Figure 7: Coefficients on Monetary Policy in Five-Year Rolling Window Estimates, Vintage-Specific Approach



Note: The figure shows the estimated coefficients on the spread (S) and lagged fed funds rate ($FF1$) using the Vintage-Specific approach with a 5-year rolling window and a horizon of 1 to 4 quarters.

[12] Had the tests turned out differently, we might have needed to formally account for multiple testing using Bonferroni bounds or some other means.

7. Summary and Conclusions

To summarize the results of this myriad of tests, I have shown that inefficiency holds in-sample, based on standard tests on the forecast errors. However, the attempt to improve on the SPF forecasts out of sample is generally not successful and sometimes makes the forecasts significantly worse. I accounted carefully for different approaches to thinking about the data-generating process, data revisions, and structural instability.

Why might in-sample results show a relationship between macroeconomic variables and forecast errors, but out-of-sample results do not? My results suggest that the time-varying coefficients on the measures of monetary policy introduce so much noise into the attempt to improve upon the forecasts that they end up being worse than the original forecasts. It may also be that forecasters do not recognize the importance of a variable for forecasting until some time passes, so there is an in-sample relationship that is not useful for forecasting for very long. Or it may take forecasters some time to adjust to structural shifts as they learn about the long run, as discussed by Farmer et al. (2024). Or, as Cukierman et al. (2020) suggest, a permanent-transitory confusion may lead to in-sample correlations, even if forecasters have rational expectations.

The structure of the forecast-improvement exercises in this paper is based on the in-sample results reported by others in the literature, cited in the Introduction. Some possible future extensions of this work include testing additional variables that might affect real GDP growth forecasts and modifying the degree of shrinkage or looking for optimal shrinkage.

The main interpretation of these results is that theories of sticky information and noisy information, and more generally informational rigidities, are not supported by the data. The literature that tested these theories was based solely on in-sample evidence. As in Eva and Winkler (2023), out-of-sample evidence provides no support for any theory of informational rigidities. While there may be periods of inefficient forecasts in sample, those inefficiencies are not exploitable,

and therefore, not sufficient to support theories that differ from rational expectations.

References

Ball, L. and Croushore, D. (2003), 'Expectations and the effects of monetary policy', *Journal of Money, Credit and Banking* **35**, 473–484.

Bianchi, F., Ludvigson, S. C. and Ma, S. (2022), 'Belief distortions and macroeconomic fluctuations', *American Economic Review* **112**(7), 2269–2315.

Bordalo, P., Gennaioli, N., Ma, Y. and Shleifer, A. (2020), 'Overreaction in macroeconomic expectations', *American Economic Review* **110**(9), 2748–2782.

Clark, T. E. and Mertens, E. (2024), Survey expectations and forecast uncertainty, *in* M. Clements and A. Galvao, eds, 'Handbook of Research Methods and Applications on Macroeconomic Forecasting', Edward Elger Publishing Ltd.

Clements, M. P. (2022), 'Forecaster efficiency, accuracy, and disagreement: Evidence using individual-level survey data', *Journal of Money, Credit and Banking* **54**(2-3), 537–568.

Coibion, O. and Gorodnichenko, Y. (2012), 'What can survey forecasts tell us about information rigidities?', *Journal of Political Economy* **120**(1), 116–159.

Croushore, D. (2011), 'Frontiers of real-time data analysis', *Journal of Economic Literature* **49**, 72–100.

Croushore, D. (2025), Improving biased forecasts in real time. University of Richmond working paper.

Croushore, D. and Stark, T. (2001), 'A real-time data set for macroeconomists', *Journal of Econometrics* **105**, 111–130.

Croushore, D. and Stark, T. (2019), 'Fifty years of the survey of professional forecasters', *Federal Reserve Bank of Philadelphia Economic Insights* pp. 1–11.

Cukierman, A., Lustenberger, T. and Meltzer, A. (2020), The permanent-transitory confusion: Implications for tests of market efficiency and for expected inflation during turbulent and tranquil times, *in* A. Arnon, W. Young and K. van der Beek, eds, 'Expectations: Theory and Applications from Historical Perspectives', Springer International Publishing, pp. 215–238.

Diebold, F. X. and Lopez, J. A. (1996), 8 forecast evaluation and combination, *in* 'Statistical Methods in Finance', Vol. 14 of *Handbook of Statistics*, Elsevier, pp. 241–268.

Diebold, F. X. and Mariano, R. S. (1995), 'Comparing predictive accuracy', *Journal of Business and Economic Statistics* **13**, 253–263.

Diebold, F. X. and Pauly, P. (1990), 'The use of prior information in forecast combination', *International Journal of Forecasting* **6**, 503–508.

Eva, K. and Winkler, F. (2023), A comprehensive empirical evaluation of biases in expectation formation. Working Paper, Federal Reserve Board.

Farmer, L. E., Nakamura, E. and Steinsson, J. (2024), 'Learning about the long run', *Journal of Political Economy* **132**(10), 3334–3377.

Harvey, D. S., Leybourne, S. J. and Newbold, P. (1997), 'Testing the equality of prediction mean squared errors', *International Journal of Forecasting* **13**, 281–291.

Koenig, E., Dolmas, S. and Piger, J. (2003), 'The use and abuse of 'real-time' data in economic forecasting', *Review of Economics and Statistics* **85**, 618–628.

Mankiw, N. G. and Reis, R. (2002), 'Sticky information versus sticky prices: A proposal to replace the new keynesian phillips curve*', *The Quarterly Journal of Economics* **117**(4), 1295–1328.

Newey, W. and West, K. (1987), 'A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix', *Econometrica* **55**, 703–708.

Rossi, B. and Sekhposyan, T. (2010), 'Have economic models' forecasting performance for us output growth and inflation changed over time, and when?', *International Journal of Forecasting* **26**(4), 808–835.

Rossi, B. and Sekhposyan, T. (2016), 'Forecast rationality tests in the presence of instabilities, with applications to federal reserve and survey forecasts', *Journal of Applied Econometrics* **31**(3), 507–532.

Rudebusch, G. D. and Williams, J. C. (2009), 'Forecasting recessions: The puzzle of the enduring power of the yield curve', *Journal of Business and Economic Statistics* **27**(4), 492–503.

Sims, C. A. (2003), 'Implications of rational inattention', *Journal of Monetary Economics* **50**(3), 665–690.

Appendix

This Appendix shows results that were removed from the body of the paper to make it more succinct.

In-Sample Results Including a Constant Term

Table 6 shows the results of the in-sample tests, including a constant term. For both clarity and simplicity in the paper, I reported in Table 3 only results that did not include a constant term. The results are similar, with a few more cases of statistically significant terms when I include a constant.

Table 6: In-Sample Results for Monetary Policy, Including a Constant Term

$e_{t,t+h} = \alpha + \beta MP_{t-1} + \epsilon_t^h$

Horizon	0	1	2	3	4	1-4	0-3
Realized Value							
initial	x x	x x	M S	S S	S S	S S	S S
first final	x x	x x	x M	x S	S S	S S	S S
first annual	x x	x M	x S	M S	S S	S S	S S
pre-benchmark	x x	x M	x S	M S	S S	S S	S S
latest-available	S S	x x	x x	S x	S S	S S	S M

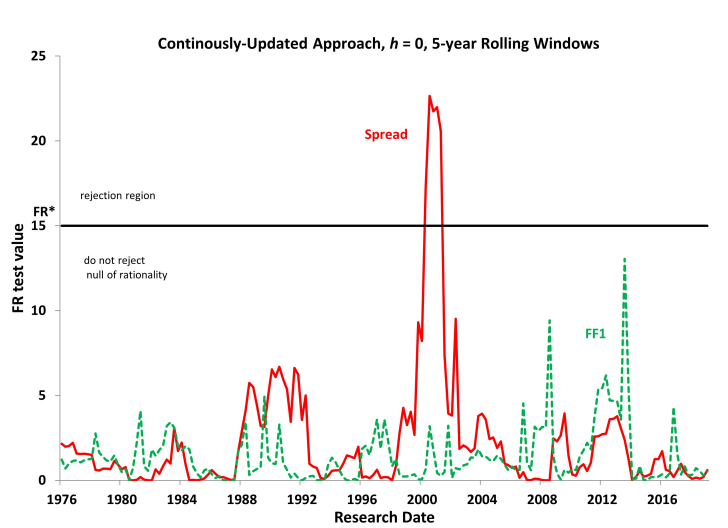
Note: Results of joint test that $\alpha = 0$ and $\beta = 0$: x means p-value > 0.10; M means $0.05 < p$-value < 0.10; S means p-value ≤ 0.05. First term: yield spread; second term: lagged change in real fed funds rate. The sample uses SPF forecasts from 1971Q1 to 2018Q4. Standard errors are adjusted following the Newey and West (1987) procedure.

In-Sample Results for Continuously-Updated Approach

In the Continuously-Updated approach, we think about a researcher assuming that forecasters at each date use the latest data from FRED or a similar macroeconomic database. The assumption is that the researcher and forecasters ignore any effects of data revisions in evaluating and forming forecasts. Follow the same rolling procedure as described in the main body of the paper. The results

of this exercise are shown in Figure 8 for the current-quarter horizon. In the discussion that follows, I show results only for the $h = 0$ quarterly horizon and the $h = 1$ to 4 annual horizon, but results for other horizons are also available.

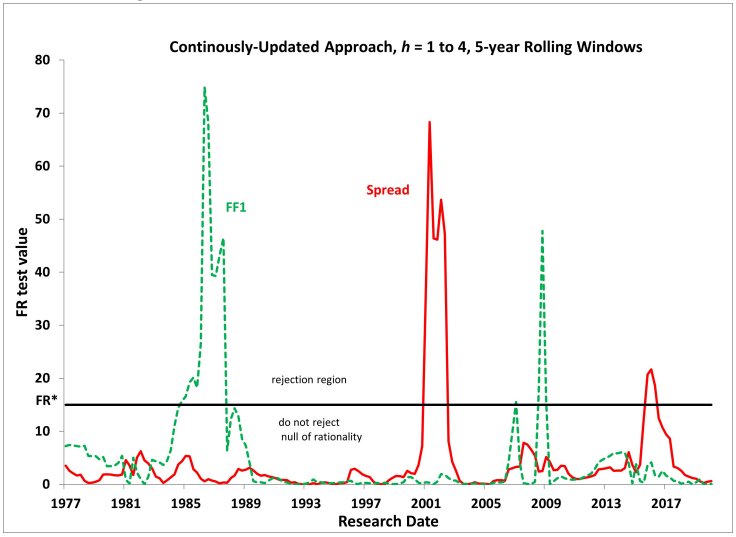
Figure 8: Forecast-Rationality Test Results, Continuously-Updated Approach, h = 0, 5-Year Rolling Windows



In Figure 8, we see just a few rejections of the null hypothesis of forecast rationality, and only with the yield spread in the in the early 2000s, and never with the $FF1$ measure. To compare with a longer-horizon forecast, Figure 9 shows the results of the forecast-rationality tests for the 1-to-4 quarter horizon. Figure 9 shows that for this longer horizon, rejections of forecast rationality occur much more frequently than for the current-quarter horizon, and are scattered across the sample.

Repeating this exercise for 10-year rolling windows leads to fewer rejections of forecast rationality than for 5-year windows, as Figures 10 and 11 show. For the current-quarter horizon, there are no rejections of forecast rationality. For the 1-to-4 quarter horizon, the only rejections come from using the $FF1$ measure of monetary policy, and just for a short period in the early 1990s. So, clearly rejections of forecast rationality depend on the horizon and the length of the

Figure 9: Forecast-Rationality Test Results, Continuously-Updated Approach, h = 1 to 4, 5-Year Rolling Windows



rolling window. There are some differences across measures of monetary policy, as well.

In-sample Results for Benchmark-Consistent Approach.

In the Benchmark-Consistent approach, we think about a researcher assuming that forecasters at each date string together pre-benchmark values of the data, with the idea that the DGP differs across benchmark revisions. The assumption is that the researcher and forecasters account for data revisions in evaluating and forming forecasts.

The forecast-rationality tests for the Benchmark-Consistent approach show a similar pattern of rejections of forecast rationality as was the case for the Continuously-Updated approach. For 5-year rolling windows, compare Figure 8 with Figure 12; and compare Figure 9 with Figure 13. For 10-year rolling windows, compare Figure 10 with Figure 14; and compare Figure 11 with Figure 15. Surprisingly, even though the approaches (Continuously-Updated and Benchmark-Consistent) are quite different, the FR tests lead to very similar

Figure 10: Forecast-Rationality Test Results, Continuously-Updated Approach, h = 0, 10-Year Rolling Windows

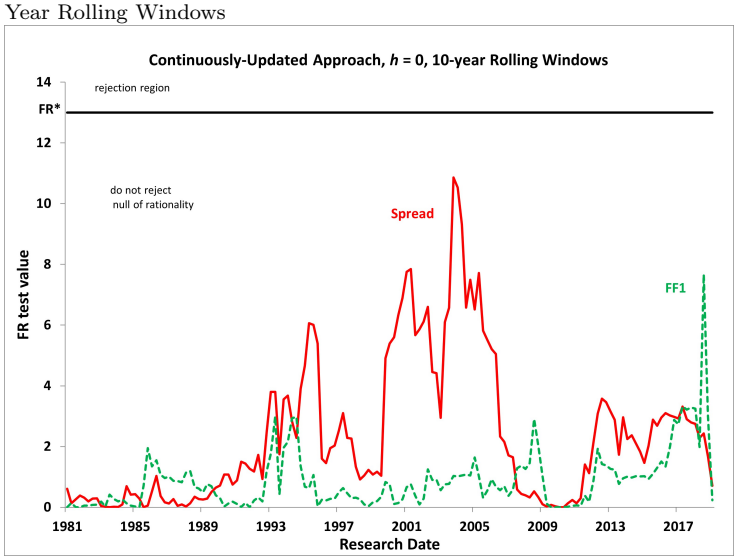
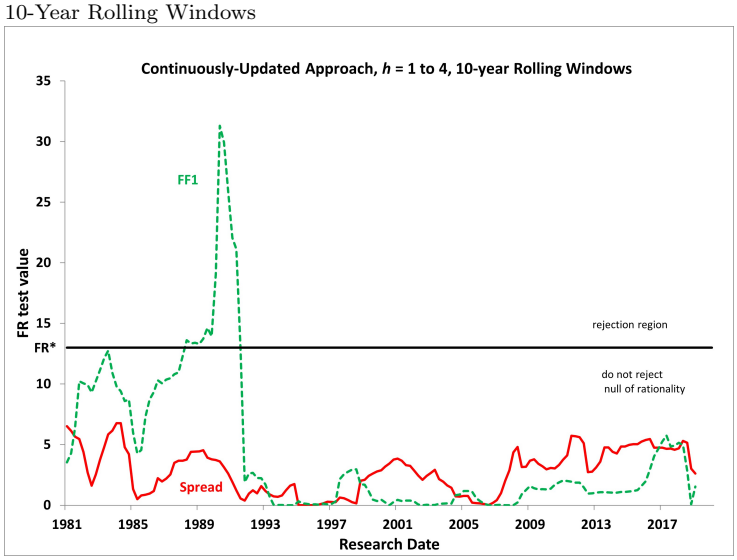
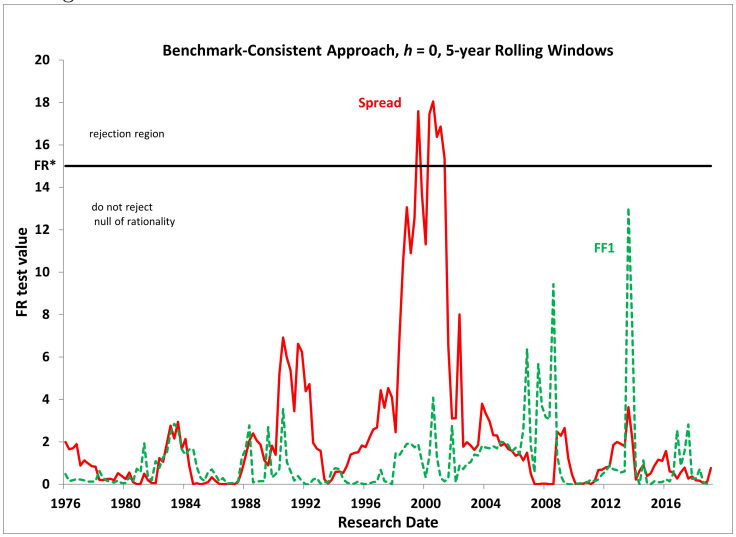


Figure 11: Forecast-Rationality Test Results, Continuously-Updated Approach, h = 1 to 4, 10-Year Rolling Windows



results.

Figure 12: Forecast-Rationality Test Results, Benchmark-Consistent Approach, h = 0, 5-Year Rolling Windows



Additional In-Sample Results for Vintage-Specific Approach. The main text showed results for the h =1 to 4 horizon. Here I show some in-sample results for the $h = 0$ horizon. For $h = 0$, Figures 16 and 17 show no rejections of the Forecast-Rationality text, so forecast improvement is not possible, as the forecasts are rational.

Results of Forecast-Improvement Exercises for All Three Approaches

Table 7 shows results of forecast-improvement exercises across the three different approaches (Continuously-Updated, Benchmark-Consistent, and Vintage-Specific). The table shows counts of $RRMSFE$s in different ranges, along with the number that are statistically significant at the 5 percent level. Looking across the approaches, it appears most likely to find improved forecasts using the Vintage-Specific approach, and less likely with the Continuously-Updated approach. The most significant forecast worsening is for Benchmark-Consistent approach, with 7 of 112 cases statistically significantly worse.

Figure 13: Forecast-Rationality Test Results, Benchmark-Consistent Approach, h = 1 to 4, 5-Year Rolling Windows

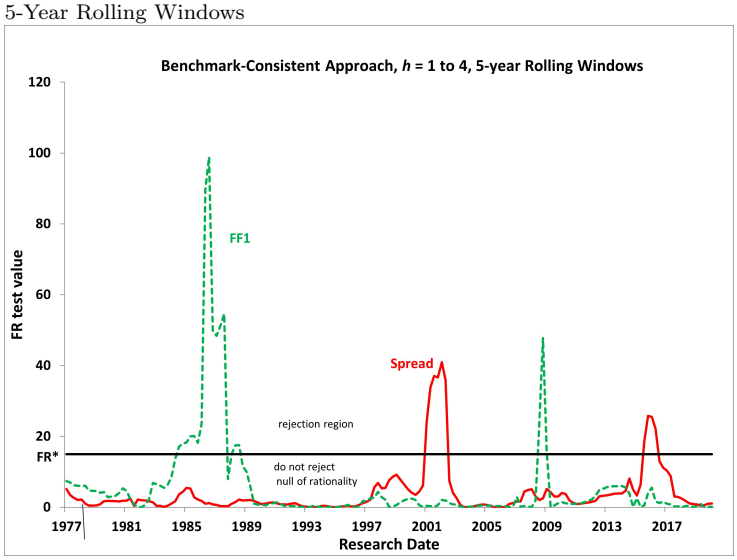


Figure 14: Forecast-Rationality Test Results, Benchmark-Consistent Approach, h = 0, 10-Year Rolling Windows

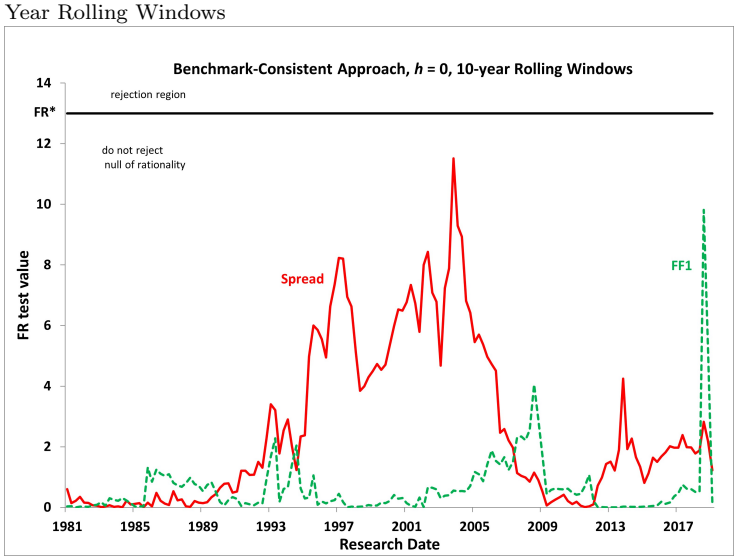


Figure 15: Forecast-Rationality Test Results, Benchmark-Consistent Approach, h = 1 to 4, 10-Year Rolling Windows

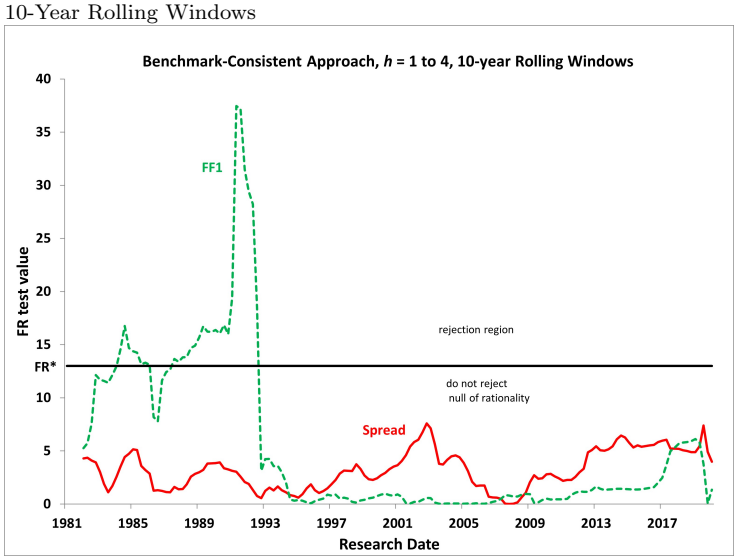


Figure 16: Forecast-Rationality Test Results, Vintage-Specific Approach, h = 0, 5-Year Rolling Windows

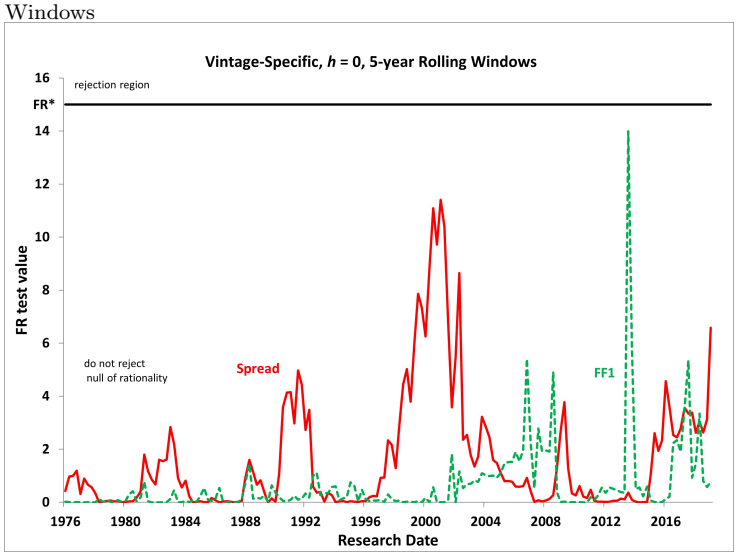


Figure 17: Forecast-Rationality Test Results, Vintage-Specific Approach, h = 0, 10-Year Rolling Windows

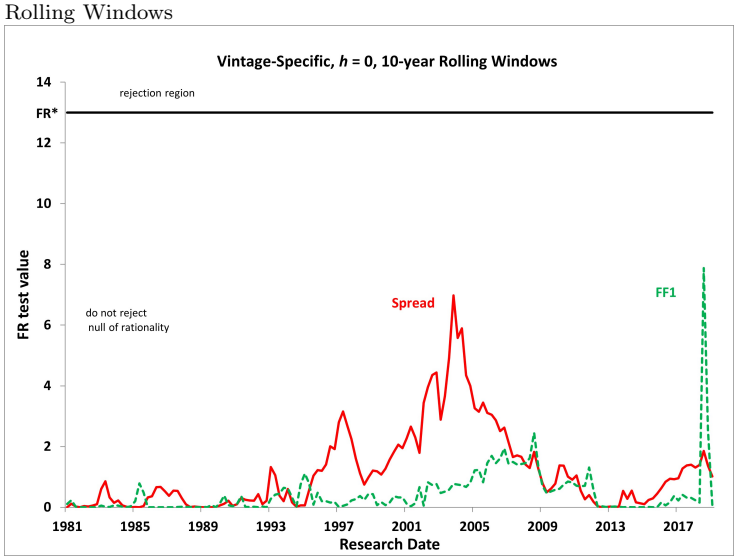


Table 7: Forecast Improvement Exercises Counts of Ranges, Across Approaches

$RRMSFE$	CU		BC		VS	
range	N	p	N	p	N	p
1.10 to ∞	7	3	6	2	2	0
1.02 to 1.10	33	2	40	5	23	4
1.00 to 1.02	44	0	34	0	46	0
1.00 exactly	22	0	22	0	28	0
0.98 to 1.00	6	0	10	0	13	0
0.90 to 0.98	0	0	0	0	0	0
0.00 to 0.90	0	0	0	0	0	0

Notes: The table shows the numbers of cases in which the relative-root-mean-squared error ($RRMSFE$) falls within the given range, and the number of cases with p-values of the Diebold-Mariano test that are less than or equal to 0.05, across all 112 permutations of 2 rolling window sizes, 2 measures of monetary policy, 7 horizons, and 4 adjustment methods for each of the 3 approaches. The sample consists of SPF forecasts made from 1971Q1 to 2018Q4.

If we repeat Table 7 but include a constant in the forecast-improvement regression, we get the results shown in Table 8. The results are similar to the case in which a constant was not included, though there are now more cases in which the forecasts are made significantly worse than the original forecasts, and also fewer cases of forecast improvement. So, dropping the constant term from the regression is beneficial to the attempt to improve on the forecasts.

Table 8: Forecast Improvement Exercises Counts of Ranges, Across Approaches, Including Constant Term

$RRMSFE$	CU		BC		VS	
range	N	p	N	p	N	p
1.10 to ∞	17	11	28	13	15	7
1.02 to 1.10	46	5	41	3	38	4
1.00 to 1.02	41	0	34	0	45	2
1.00 exactly	4	0	4	0	10	0
0.98 to 1.00	4	0	5	0	4	0
0.90 to 0.98	0	0	0	0	0	0
0.00 to 0.90	0	0	0	0	0	0

Notes: The table shows the numbers of cases in which the relative-root-mean-squared error ($RRMSFE$) falls within the given range, and the number of cases with p-values of the Diebold-Mariano test that are less than or equal to 0.05, across all 112 permutations of 2 rolling window sizes, 2 measures of monetary policy, 7 horizons, and 4 adjustment methods for each of the 3 approaches, when a constant term in included in the regression. The sample consists of SPF forecasts made from 1971Q1 to 2018Q4.