

Real-Time Uncertainty in Estimating Bias in Macroeconomic Forecasts

By DEAN CROUSHORE*

September 7, 2023

In this paper, we examine how tests for bias in expectations of output growth and inflation, measured using the Survey of Professional Forecasters, have changed over time. The results of bias tests are found to depend on the subsample in question, as well as what concept is used to measure the actual value of a macroeconomic variable and the empirical technique used. We find that although bias appears in sub-samples, it is difficult, though not impossible, to improve the forecasts by exploiting the measured bias in real time.

JEL: E37, E17

Keywords: real-time data, output forecasts, inflation forecasts

* University of Richmond, Robins School of Business, dcrousho@richmond.edu. I thank participants at seminars and conferences, including Workshop on Real-Time Data Analysis, Southern Economic Association meetings, Society for Nonlinear Dynamics and Econometrics, Society for Computational Economics, West Virginia University, CIREQ, Workshop in Macroeconomic Research at Liberal Arts Colleges, University of Richmond, University of Alabama, and North Carolina State University.

Economists are constantly looking for stylized facts. One of the most important stylized facts that economists have tried to establish (or disprove) is that forecasts are rational. The theory of rational expectations depends on it, yet the evidence is mixed. Whether a set of forecasts is found to be rational or not seems to depend on many things, including the sample, the source of data on the expectations being examined, and the empirical technique used to investigate rationality.

Early papers in the rational-expectations literature used surveys of expectations, such as the Livingston Survey and the Survey of Professional Forecasters, to test whether the forecasts made by professional forecasters were consistent with the theory. A number of the tests in the 1970s and 1980s cast doubt on the rationality of the forecasts, with notable results by Su and Su (1975) and Zarnowitz (1985). But later results, such as Croushore (2010), find no bias over a longer sample.

Croushore (2010) found substantial instability across subsamples in evaluations of survey forecasts of inflation in a manner similar to that found by Giacomini and Rossi (2010) for model forecasts of exchange rates. No global stylized facts appear to hold. Forecasters go through periods in which they forecast well, then there is a deterioration of the forecasts, and then they respond to their errors and improve their models, leading to lower forecast errors again. This pattern may explain why Stock and Watson (2003) find that many variables lose their predictive power as leading indicators. Perhaps parameters are changing in economic models, as Rossi (2006) suggests for models of exchange rates.

The motivating question of this paper is: does the concept chosen to represent the realized value or “actual” matter, along with the subsample? The term “actual” is in quotes because it can have many meanings. In this case, it refers to the idea that data are revised; therefore it may not be clear which concept forecasters are targeting. If data revisions are not forecastable, forecasters would generate the same forecasts, whether they are trying to forecast the initial release of a macroeconomic variable, or the annual revised value, or some final,

revised version. Because data revisions persist through time, data are never final. Researchers must choose between many different concepts of actual data.

The central message of this paper is consistent with the work of Rossi (2006) and Stock and Watson (2003). Not only is the performance of different types of forecasts unstable, but the timing of that instability depends on the data vintage being used in the analysis. The overall conclusion is that we are unlikely to find stylized facts about rational expectations as measured by economic forecasts. In fact, we find that although bias appears in sub-samples, it is difficult to improve the forecasts by exploiting the measured bias in real time.¹

We begin by describing the data on forecasts and the real-time data used to evaluate the forecasts in section I. In section II, we show the results of unbiasedness tests and how the outcomes of such tests depends on where the tester stands in time and when the tester's sample begins. Section III describes tests to see if the sub-sample bias found in the previous section can be exploited in real time to improve the forecasts, using initially released data and using latest-available data. Finally, section IV interprets the results and provides conclusions.

I. Data

In this paper, we study two different variables: the growth rate of real output and the inflation rate as measured by the growth rate of the GDP price index. The complication for both variables is that, because they are revised over time, these data revisions may pose difficulties in evaluating the accuracy of the forecasts, as suggested by Croushore (2011). We handle this complication by using the real-time data set of Croushore and Stark (2001). Data are available for both variables from data vintages beginning in the third quarter of 1965, when quarterly real output was reported for the first time on a regular basis by the U.S. Bureau of

¹A recent paper that is complementary to this one using similar methods and evaluates many more variables is Eva and Winkler (2023). That paper places less emphasis on sub-sample stability and alternative measures of actuals.

Economic Analysis.²

To study the ability of forecasters to provide accurate forecasts, we use the Survey of Professional Forecasters (SPF), which records the forecasts of a large number of private-sector forecasters.³ The literature studying the SPF forecasts has found that the SPF forecasts outperform macroeconomic models, even fairly sophisticated ones, as shown by Ang, Bekaert and Wei (2007). The SPF has also been found to influence household expectations, as shown by Carroll (2003).

While some arguments can be made that testing rational expectations is best done by examining the forecasts of individual forecasters,⁴ a more compelling argument is that the most accurate forecasts are provided by taking the mean across the forecasters, as illustrated by Aiolfi, Capistran and Timmermann (2011). An additional problem with using the forecasts of individual forecasters is that the SPF survey has many missing observations, so finding statistically significant differences across individual forecasters is problematic. Data on mean and median forecasts of output and inflation are reported in the SPF beginning with the fourth quarter of 1968.⁵ However, the forecasts in the early years of the survey were not reported to enough significant digits, and four-quarter-ahead forecasts were sometimes not reported in the early years of the survey. To avoid these problems, we begin our analysis using surveys beginning from the first quarter of 1971.

There are many horizons for the SPF, and in this paper we choose to focus on the longest forecasting horizon that is consistently available in the survey, which is the average growth rate of output (or average inflation rate) over the next year (four quarters), labeled “One-year ahead” forecasts. The one-year-ahead forecast is subject to less noise and presumably more economic causes than would be the case for studying the forecasts for a particular quarterly horizon.

²See the documentation on the Federal Reserve Bank of Philadelphia Real-Time Data Set for Macroeconomists at www.philadelphiafed.org/research-and-data/real-time-center/.

³Details on the SPF can be found in Croushore and Stark (2019).

⁴See Keane and Runkle (1990).

⁵The mean and median across forecasters are almost identical in the SPF. This paper reports results based on the mean forecast but all tests reported in the paper have also been done with median forecasts, with no material differences in results.

We begin by looking at the forecasts and forecast errors in Figure 1 for output growth and Figure 2 for inflation. The figures are based on using the initial data release as actual; of course, other concepts of actual could be used.⁶ They show some periods of persistent forecast errors, especially in the 1970s, but also at other times. However, this persistence is overstated by the figures because of the overlapping-observations problem: we are observing the forecasts quarterly, but they are four quarters ahead from the forecast date, and five quarters ahead of the last observation in the forecasters' data set. The overlapping-observations problem leads to the correlation of forecast errors. In our empirical work, we will use standard techniques to overcome this problem, adjusting the variance-covariance matrix using techniques developed by Newey and West (1987).

If revisions to the data were small and white noise, the use of different concepts for actual output growth and the actual inflation rate would be inconsequential. But the literature on real-time data analysis (see Croushore (2011)) suggests that the revisions are neither small nor innocuous. We consider six different concepts for actual output and inflation: (1) the initial release, which comes out at the end of the first month following the end of a quarter; (2) the first revision, which occurs one month after the initial release; (3) the first-final release, also called the second revision, which comes out at the end of the third month following the end of a quarter;⁷ (4) the annual release, which is usually produced each year at the end of July and usually includes revisions to data from the prior three calendar years; (5) the pre-benchmark release, which is the last release of the data prior to a benchmark revision that makes major changes in the data construction process; and (6) the last release of July 2023, which is the most recent vintage of the data at the time of writing this paper, which incorporates many benchmark revisions.⁸ In years in which a benchmark release occurs, such

⁶More discussion of these other concepts occurs later in this paper.

⁷This is the value for actuals that is used often in the literature, for example, by Romer and Romer (2000) and Rudebusch and Williams (2009).

⁸We use the date July 2023 in this paper; it corresponds to the vintage of August 2023 in the Philadelphia Fed's Real-Time Data Set for Macroeconomists (RTDSM), the timing of which is in the middle of the month. So, the data released at the end of July 2023 are recorded in the August vintage

as 2003, there is often no annual revision, so we take the benchmark release of the data as the annual release. The pre-benchmark release is an important concept because it shows the last data following a consistent methodology. For example, before 1996, macroeconomic forecasters all based their forecasts on fixed-weighted GDP. But in early 1996, when the government introduced chain-weighted GDP, the entire past history of GDP changed substantially. A forecaster who made a forecast of GDP growth in 1994 would not have produced forecasts of chain-weighted GDP, so it seems appropriate to compare those forecasts to the last release of the data containing fixed-weighted GDP. For complete details on these concepts and the revision process, see Croushore (2011).⁹

With many vintages of data and alternative concepts of actuals, it is useful to set some notation that will help clarify the various experiments in which we engage. The variables we use follow the same pattern of data releases and revisions.

In general terms, we use a subscript to denote the quarter for which the data apply and a superscript to denote the date of the vintage, where a subscript has two terms: the quarter of the vintage, and the month. For example, the last observations in our sample were released at the end of July 2023 and the last quarter for which data exists is the second quarter of 2023. So, the value of variable X for that date in its initial release is denoted as:

$$X_{2023Q2}^{2023Q3,1}.$$

We will denote all the data in the last release (July 2023), which contains data from 1947Q1 to 2023Q2, as:¹⁰

$$X^{last} = \{X_{1947Q1}^{2023Q3,1}, X_{1947Q2}^{2023Q3,1}, X_{1947Q3}^{2023Q3,1}, \dots, X_{2022Q4}^{2023Q3,1}, X_{2023Q1}^{2023Q3,1}, X_{2023Q2}^{2023Q3,1}\}.$$

of the RTDSM.

⁹The Appendix shows summary statistics for the forecast errors using these six different concepts, as well as the dates of both annual revisions and benchmark revisions.

¹⁰The data come from the Real-Time Data Set for Macroeconomists (RTDSM), the vintages of which are dated mid-month. So, the data released at the end of July 2023 are called the vintage of 2023M8 (August 2023) in the RTDSM.

Similarly, any other vintage of data can be described as:

$$X^{Q,M} = \{X_{1947Q1}^{Q,M}, X_{1947Q2}^{Q,M}, \dots, X_{Q-1}^{Q,M}\}.$$

For example, the data release at the end of January 1999 is:

$$X^{1999Q1,1} = \{X_{1947Q1}^{1999Q1,1}, X_{1947Q2}^{1999Q1,1}, \dots, X_{1998Q4}^{1999Q1,1}\}.$$

Thus, based on our earlier definition, $X^{last} = X^{2023Q3,1}$.

The first regular monthly release of quarterly GDP data occurred at the end of October 1965 and the last observation in that release was for 1965Q2. Almost always,¹¹ the first release for output and the price level occurred in the first month of the following quarter, so we denote a collection of all the initial releases as:

$$X^{initial} = \{X_{1965Q2}^{1965Q3,1}, X_{1965Q3}^{1965Q4,1}, X_{1965Q4}^{1966Q1,1}, \dots, X_{2022Q4}^{2023Q1,1}, X_{2023Q1}^{2023Q2,1}, X_{2023Q2}^{2023Q3,1}\}.$$

The first-revision actuals are similar to the initial actuals but use the data vintage from the second month of each quarter. Similarly, the first-final actuals use the data vintage from the third month of each quarter.

Annual revisions usually occur every year at the end of July, with some exceptions (including in July 2023), which we note in the Appendix. Collecting all the first annual revisions gives us the following vector:

¹¹The exception was the first release of 1995Q4, which was delayed because of the federal government shutdown.

$$\begin{aligned}
X^{annual} = & \{X_{1965Q2}^{1966Q3,1}, X_{1965Q3}^{1966Q3,1}, X_{1965Q4}^{1966Q3,1}, \\
& X_{1966Q1}^{1967Q3,1}, X_{1966Q2}^{1967Q3,1}, X_{1966Q3}^{1967Q3,1}, X_{1966Q4}^{1967Q3,1}, \\
& \dots, \\
& X_{2020Q1}^{2021Q3,1}, X_{2020Q2}^{2021Q3,1}, X_{2020Q3}^{2021Q3,1}, X_{2020Q4}^{2021Q3,1}, \\
& X_{2021Q1}^{2022Q3,1}, X_{2021Q2}^{2022Q3,1}, X_{2021Q3}^{2022Q3,1}, X_{2021Q4}^{2022Q3,1}\}.
\end{aligned}$$

The pre-benchmark values are more difficult to generate, as their pattern is irregular. Dates for the pre-benchmark vintages are given in the Appendix. The first benchmark revision was in late January 1976, so the pre-benchmark values came from the December 1975 (1975Q4,3) vintage. If there has not yet been a benchmark revision for some observations, we use the last vintage available. The overall vector looks like:

$$\begin{aligned}
X^{pre-benchmark} = & \{X_{1965Q2}^{1975Q4,3}, X_{1965Q3}^{1975Q4,3}, X_{1965Q4}^{1975Q4,3}, \\
& X_{1966Q1}^{1975Q4,3}, X_{1966Q2}^{1975Q4,3}, X_{1966Q3}^{1975Q4,3}, X_{1966Q4}^{1975Q4,3}, \\
& \dots, \\
& X_{2020Q1}^{2022Q3,1}, X_{2020Q2}^{2022Q3,1}, X_{2020Q3}^{2022Q3,1}, X_{2020Q4}^{2022Q3,1}, \\
& X_{2021Q1}^{2022Q3,1}, X_{2021Q2}^{2022Q3,1}, X_{2021Q3}^{2022Q3,1}, X_{2021Q4}^{2022Q3,1}\}.
\end{aligned}$$

In evaluating forecasts, researchers have used different measures of actuals. Most common in early literature was the use of the latest available vintage of data available to the researcher. Indeed, assuming that data revisions get us closer to the truth over time, this seems reasonable. However, the data have undergone numerous conceptual revisions over time, which changed the definition of output. For example, it is difficult to imagine that a forecaster in 1971

would account for the future change of the output concept to include intellectual property products, which caused GDP for most periods to be revised up after the benchmark revision of July 2013, when the concept of intellectual property products was introduced. For that reason, some researchers, such as Zarnowitz (1985), prefer to use a concept like the pre-benchmark release, while others, such as Croushore (2011) focus on the first annual revision or first-final (third) release, or even the initial release. The real-time literature has shown that some empirical results are sensitive to the choice of actual.

In addition to the choice of actuals, different vintages may need to be used to get an accurate portrayal of the data-generating process. Most prominently, Kishor and Koenig (2012) show that the correct relationship across vintages may depend on the vintage concept; for example, the sequence of initial releases may have a separate data-generating process than later releases of the data. We explore different possibilities in this paper.

II. Results

A. Tests for unbiasedness over full sample

In this paper, our focus is on tests for the unbiasedness of forecasts. In the literature on forecast bias, the standard test is the Mincer and Zarnowitz (1969) test, which regresses actual values on forecasts. However, the Mincer-Zarnowitz test may be inaccurate in small samples, as Mankiw and Shapiro (1986) show. Because we are using small samples, and because some of the tests we perform will be sensitive to parameter uncertainty, we will modify the test for unbiasedness to a simpler version, which tests whether the forecast error has a mean of zero.¹²

We run the zero-mean-forecast-error test for both output growth and inflation, using all six versions of actuals and the data vector X^{last} , where X is either

¹²We follow most of the forecasting literature in testing for bias under the assumption of a loss function for which bias is undesirable. A few papers, such as Elliott, Komunjer and Timmermann (2008), allow for the possibility that the loss function of forecasters may be asymmetric, which implies that bias in the forecasts may be optimal.

output growth or inflation. The forecast error is the actual value (one of the six possibilities) minus the forecast.

The results of this exercise are shown in Table 1. In each case, we show the mean forecast error, the p -value from the t -test for whether the mean forecast error is significantly different from zero, and the standard error. Table 1 shows that for all versions of actuals and for both variables, we never reject the null hypothesis of zero-mean forecast error, with most p -values well above 0.05.¹³

As Figures 1 and 2 suggest, however, the COVID period represented a huge shock that forecasters could not have possibly forecast well, so perhaps the results in Table 1 are distorted by COVID. To test that, we rerun the bias tests so that they end before the COVID period, as shown in Table 2. The results are consistent with those in Table 1, with no rejection of the null hypothesis of zero-mean-forecast errors. But notice that the mean errors, p -values, and standard errors all differ from the period that includes COVID. For the remainder of this paper, we will focus on the pre-COVID period but we did the analysis for the period including COVID with no major differences in the results.

B. Tests for unbiasedness in sub-samples

Croushore (2010) shows that results like the tests for bias shown in Tables 1 and 2 tend to be fragile: they change dramatically depending on the precise beginning and ending dates of the sample. One way to investigate this is to consider how researchers might have perceived the bias at various points in (vintage) time. Suppose a researcher had run the zero-mean test in the second quarter of 1979, with data and one-year-ahead forecasts made from 1971Q1 to 1978Q1, that is, using the data set $X^{1979Q2,1}$. What conclusion about bias would she have drawn? We can ask the same question for a researcher standing at any date between 1979Q3 and 2023Q2. But doing so is a bit difficult because we must be careful

¹³The Appendix shows additional tables containing other statistics: root-mean-squared errors and mean absolute errors.

to consider the exact information set a researcher would have at each date. For example, a researcher would most likely use the latest-available vintage at each date to evaluate the past forecasts; so the researcher might use latest-available actuals at each date. But, of course, a researcher standing in the second quarter of 1978 would have had a very different version of the latest-available data than the August 2023 vintage that is the last one we use in this paper. So, we can collect a sequence of latest-available data sets at each date. We call the date at which a researcher would observe those data as the “research date.” Of course, such a data set uses data that have been subject to benchmark revisions, which may be less than desirable.

A second way of thinking about what a researcher might do, is to run the unbiasedness regression using just initial actuals, as Koenig, Dolmas and Piger (2003) suggest, based on the idea that the same actual concept is the proper object in the data-generating process. They call the procedure of using latest-available actuals each period, as described in the previous paragraph, as using “end-of-sample” data (the *EOS* method). They suggest instead using initial data releases as actuals for every period, which they term the “real-time vintage” approach (*RTV*).¹⁴ So, now we still consider a researcher standing at different points in time, beginning in the second quarter of 1978, but using the *RTV* approach. In the second quarter of 1978, the research would use the $X^{initial}$ data set and one-year-ahead forecasts from 1971Q1 to 1976Q4. The researcher would test for bias on that data set, then roll forward and repeat the exercise once each quarter after that (when new initial data were released) until finally evaluating the one-year-ahead forecasts made in 2022Q2 using the data set in 2023Q3.

The results of these *EOS* and *RTV* exercises are shown in Figures 3 (for output) and 4 (for inflation). The top panel in each figure shows the p -values for the test of unbiasedness, while the lower panel shows the estimate of the bias, in percentage

¹⁴Other concepts of actual could also be used in this exercise but do not show major differences from the use of the initial values as actuals.

points. The blue dotted lines are based on the *EOS* approach, while the red solid lines are based on the *RTV* approach.

The two approaches lead to the same general outcomes. There are not many significant rejections of the null of no bias—only for a few cases for inflation when the sample ends in 1979 or in the early 1980s. Of course, this is the period in which the rational-expectations hypothesis was gaining popularity and being tested. Numerous researchers found this bias for inflation forecasts and argued against rationality in the forecasts.¹⁵ However, as we can see in the figures, those rejections were short-lived, and as the sample of data increased, the null of no bias was no longer rejected.

Notice that the estimated bias changes as the research date gets later. In short samples, with research dates in the late 1970s and early 1980s, the bias for output growth is often estimated at around -0.5 to -1.0 percentage points. But as the research date gets later, and thus the sample gets longer, the estimated bias gets closer to zero. Similarly, for inflation, the estimated bias is around $+1.5$ percentage points with early research dates, then gets very close to zero as the sample size increases with later research dates. Thus, the results in the literature of finding bias in papers written in the 1970s and 1980s may not be surprising; while papers written in the last 20 years usually find no bias.

An alternative way of looking at this issue of subsample stability is to consider it from another point of view: what would have happened if the survey had come into existence later?¹⁶ So, consider subsamples that end in 2023Q2, but begin at various dates after 1971Q1. Figures 5 and 6 shows the results of this exercise. In this case, we are taking the 2023Q2 end date as fixed and considering the estimated bias that a researcher would calculate, using both the *EOS* and *RTV* methods, for different starting dates. For *RTV* results, we consider many different

¹⁵In the last section, we will discuss these papers in more detail.

¹⁶Different surveys of forecasters began at different dates, and researchers usually use the sample period for which forecasts are available. For example, the Blue Chip Economic Indicator survey began in 1976, while the Wall Street Journal survey began in 1986. So, the experiment in this section can shed light on why researchers find differences in bias tests across surveys.

alternatives to use as actuals, but in this paper we'll report just those for initial and annual actuals.¹⁷

In these figures, we can see that we reject the null of no bias in many additional subsamples, especially for output growth. This is consistent with the results in the literature on testing rational expectations. Notice also that for inflation, the subsample periods with rejections of the null hypothesis vary significantly across different versions of the variable used as actual, with first final actuals showing bias for many more subsamples than the other actuals. For output growth, rejections of the null hypothesis occur only for samples starting after the late 1990s, suggesting that the results are quite sensitive to the precise choice of sample period. Note also that the estimated bias for both output growth and inflation changes significantly with changes in the sample beginning date.

C. Tests with rolling windows

These results suggest that although our full-sample results led to no rejections of the null of unbiasedness, there is a lot of variation in bias over time. One way to capture this variation is to consider rolling sample windows. That is, suppose we measure the bias at each date for the previous 5 or 10 years, instead of going back to the start of the survey. So, we perform a similar zero-mean forecast-error test as before, but for rolling 5-year and 10-year windows. The p -values for this exercise are shown in Figures 7 (output growth) and 8 (inflation). We have used both the *EOS* method and the *RTV* method with initial actuals to test for bias in the rolling windows.

For both sets of rolling windows, we observe significant variation in the outcomes of the test for unbiasedness. The results depend both on the ending date of each subsample and on the method (*EOS* or *RTV*) used. For each variable, for both rolling-window sizes, there is periodic evidence of persistent bias in the

¹⁷We have results for first-final and pre-benchmark actuals, as well, but the patterns for those were similar to either initial or annual actuals.

forecasts.

D. Tests for optimality in rolling windows

A superior method of testing for the optimality of forecasts in rolling samples was developed by Rossi and Sekhposyan (2016). They develop a general test for forecast optimality that is robust to the presence of instabilities, which are suggested by this paper's results. The test allows for instabilities that cause breaks in the data or tests of unbiasedness or efficiency. We implement their test here using both 5-year rolling windows and 10-year rolling windows. The advantage of their test is that it accounts for sequential testing bias.

Figures 9 and 10 show the results of the fluctuation optimality tests. In Figure 9, we examine forecasts of output growth, with part a showing 5-year rolling windows and part b showing 10-year rolling windows. Similarly, Figure 10 examines inflation forecasts, again with part a showing 5-year rolling windows and part b showing 10-year rolling windows. In each figure, the horizontal line shows the critical value of the test at the 5 percent significance level from Rossi and Sekhposyan (2016), Table 1b.

The results for output growth in Figure 9 are mixed. For the 5-year rolling forecasts shown in panel a, there is evidence of nonoptimality in the SPF forecasts because samples ending in the second half of the 1990s and early 2000s, in 2007, and from 2014 to 2016, show test values exceeding the critical value. The interpretation of the test is that there is evidence against forecast optimality when any test value exceeds the critical value in the entire out-of-sample forecasting period. For the 10-year rolling forecasts shown in panel b, however, the test value never exceeds the critical value, suggesting that the output growth forecasts from the SPF are not suboptimal.

For inflation forecasts, the fluctuation optimality tests are much worse, with test statistics that are very large compared with the critical value. In Figure 10, illustrating inflation forecasts with 5-year rolling windows in panel a, subsamples

that end in the second half of the 1980s and most of the 1990s show very large test values far above the critical value. In panel b, the inflation forecasts with 10-year rolling windows from 1990 to 2003 show rejections. Thus, consistent with our other evidence, forecasts of inflation seem subject to biases that cause the forecasts to be suboptimal.

III. Forecast-Improvement Exercises

A problem in the literature on forecast evaluation is that many researchers find bias or inefficiency in-sample, but that bias cannot be exploited out of sample. We would like to be able to use the results of the bias tests to show that, in real time, a better forecast could have been constructed. In the early rational-expectations literature, the bias that was found in the forecasts was clear, and the prescription for researchers and policymakers was that they could improve on published forecasts by adjusting the forecasts by the amount of the bias.

A. Improving forecasts using initial releases

To improve the forecasts, given that the *RTV* method showed bias in numerous sub-samples, we estimate the bias using the initial release of the data to determine the mean forecast error, then create a new and improved forecast from the survey forecast:

$$(1) \quad \hat{\pi}_t = \hat{\alpha} + \pi_t^f,$$

where $\hat{\pi}_t$ is the new and improved forecast, $\hat{\alpha}$ is the estimated bias, and π_t^f is the published survey forecast. So, the real question is: can someone estimate the bias and make a better forecast? In this exercise, we begin by using the *RTV* method with the initial release of the data to determine the forecast error. It might be possible to use a later release of the data as well, but that creates problems in a real-time forecast-improvement exercise because concepts other than the initial

release mean longer lags in data availability. For example, using pre-benchmark data as actuals to determine the forecast error means that in real time there might be five years that pass before you get any new observations to use.

The results of this exercise are shown in Table 3. The rows of the tables show alternative experiments, described below. The first column of numbers shows the relative-root-mean-squared forecast error (*RRMSFE*) for estimating the bias using 5-year rolling windows and Equation (1), where *RRMSFE* is the *RMSFE* of the improved forecast divided by the *RMSFE* of the original survey. Thus, an *RRMSFE* less than one means that estimating the bias and using Equation (1) leads to a lower *RMSFE* and an improved forecast; an *RRMSFE* greater than one means that the attempt to improve the forecast failed. The second column of numbers shows the *p*-value for the test of a significant difference in *RMSFEs*, based on the Diebold and Mariano (1995) test.¹⁸ The next two columns repeat this exercise for 10-year rolling windows.

The first row in Table 3 labeled “Adjust every period” for both output growth and inflation shows the results of the basic experiment in which we use Equation (1) to attempt to improve on the survey forecasts based on the estimated bias each period. In almost every case, the forecasts are worse, as the *RRMSFE* is greater than one, so the *RMSFE* is higher than for the original survey. However, the *p*-values are all above 0.05, meaning that the difference in *RMSFEs* is not statistically significant.

Part of the reason for the poor performance of these attempts at forecast improvement is that we are trying to use the estimated bias even in periods when the bias is not statistically significant. However, more likely someone estimating bias in real time would adjust the forecast using Equation (1) only if the bias was statistically significantly different from zero. So, suppose we follow this strategy. We will apply Equation (1) only in periods when the *p*-value in Figures 7 or 8

¹⁸The test is subject to the caveat that with real-time data, subject to revisions, the test is not completely valid, per Clark and McCracken (2009), for which there is not yet a satisfactory solution.

is below 0.05, or when the fluctuation optimality test is rejected, as shown in Figures 9 or 10.

The results of this exercise are shown in the rows for each variable in Table 3, labeled “Adjust when $\rho < 0.05$ ” and “Adjust when FO test rejects.” Compared with the results in the first row, the results here suggest some ability to improve upon the SPF forecasts. For output growth this method is not fruitful, as the *FO* test is almost never rejected anyway. But for inflation in both 5-year and 10-year rolling windows, the *RRMSFE* is less than one when using Equation (1) to improve the forecast when the p -value is below 0.05 or when the *FO* test shows rejection. However, in no case is the difference statistically significant. So, although it is possible to use Equation (1) to improve the forecast in some cases, the forecast improvement is modest.

One final possibility is to recognize that the bias is estimated with error, so it makes sense to use shrinkage methods to reduce the error introduced by parameter estimation. Suppose we adjust for bias, but only adjust for the bias by a factor of one-half:

$$(2) \quad \hat{\pi}_t = (0.5 \times \hat{\alpha}) + \pi_t^f,$$

Using Equation (2) instead of Equation (1), we get the results shown in Table 4. We can use shrinkage, adjusting every period, or only when the p -value is below 0.05, or when the FO test shows rejection.

The results show that shrinkage generally helps. In two cases for inflation forecasts, the *RMSFE* improvement is statistically significant at the 5-percent level. Although we could search for the optimal degree of shrinkage, this would violate the concept of a researcher being able to adjust for the bias in real time.

B. Using the *EOS* method to improve forecasts

The exercises reported in Tables 3 and 4 use the initial release of the data at each date to estimate the bias using the *RTV* method. However, a more common procedure in practice is for a researcher to use the latest-available data at a given date to estimate the bias and try to use it to improve the forecasts, rather than using the initial data that were released, which is the *EOS* method. That is, suppose that at each research date, a researcher gathers the most recent data from a current database and uses those data to estimate the bias, ignoring real-time data-revision issues altogether. Based on those estimates, suppose the researcher were to use Equation (1) or (2) to improve upon the survey results, as before. The results of this exercise are shown in Tables 5 and 6, using initial actuals.

The results show that about half of all cases show an increase in *RMSFE* and about half show a decrease. In only one case (with rolling 5-year windows using shrinkage to forecast inflation when the *FO* test rejects) is there a statistically significant improvement in *RMSFE* at the 5-percent level.

The results shown in Table 5 and 6 are based on using initial actuals. However, we could have used other versions of actuals. The results do not seem terribly sensitive to the choice of actual that is used in this exercise. Tables 7 and 8 show that the *RRMSFE* does not change very much when using pre-benchmark actuals in place of initial actuals when evaluating the outcomes of the forecast-improvement exercises. Once again, a statistically significant reduction of *RMSFE* occurs for inflation forecasts using shrinkage when the *FO* test rejects.

Overall, testing numerous methods to improve on the forecasts shows that forecast improvement is not easy, though not impossible.

IV. Interpretation and conclusions

Our analysis of the variation in results across subsamples and alternate versions of actuals can explain many of the results about bias in survey forecasts of inflation and output growth in the literature. The earliest report of bias in the SPF data is that of Su and Su (1975), who found bias in the inflation forecasts in the very early years of the survey from 1968 to 1973. This result is consistent with our Figure 4, which shows bias for latest-available actuals in the early years of the survey's existence, although our sample is a bit different from theirs. Zarnowitz (1985) had a much bigger sample than did Su and Su, from 1968 to 1979, and rejected unbiasedness for SPF inflation forecasts at all horizons using pre-benchmark data. That result is perfectly consistent with our result in Figure 4, because he ran his tests in the early period during the one subperiod where unbiasedness is rejected. However, Hafer and Hein (1985) found no bias in SPF inflation forecasts from 1970 to 1984 and in subperiods from 1975 to 1979 and 1980 to 1984, but bias in the subperiod of 1970 to 1974, which is consistent with the erratic nature of rejections of unbiasedness in the 5-year windows shown in Figure 8. Bonham and Dacy (1991) found support for unbiasedness in SPF inflation forecasts from 1970 to 1984 using latest-available data, also consistent with Figure 4. Romer and Romer (2000) used the second revision of the data and the Mincer-Zarnowitz test to evaluate SPF inflation forecasts from 1968 to 1991, finding that they are unbiased, which is again consistent with Figure 4. Mankiw, Reis and Wolfers (2003) found no evidence of bias in SPF inflation forecasts from 1969 to 2002, where the lack of bias is suggested by Figure 4. So, all the results seem to mesh well with the results in this paper.

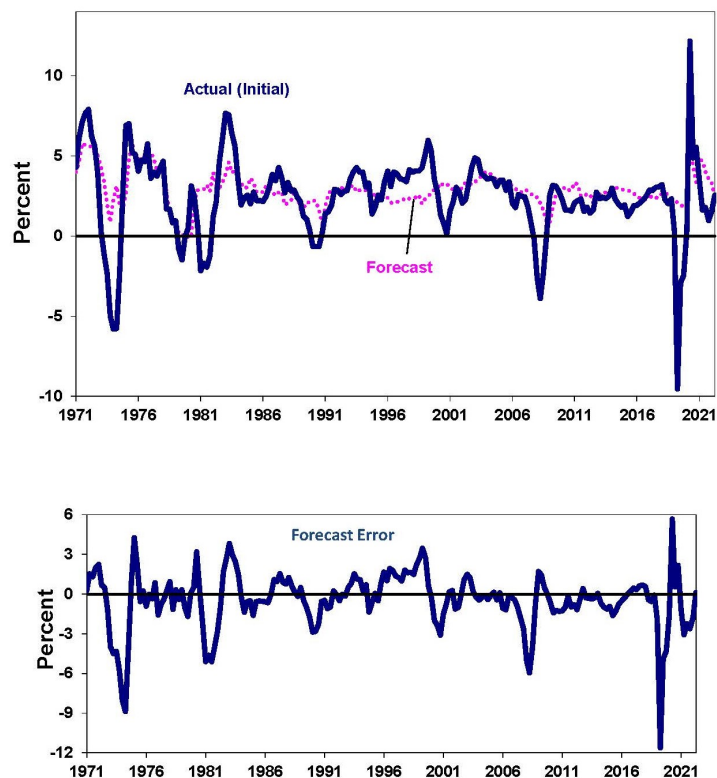
The conclusions of this paper are that (1) there are no simple stylized facts about bias in survey forecasts of output growth and inflation; (2) many subsamples of survey data show evidence of bias, even though no bias is apparent in the full sample; (3) it may be possible to improve on the survey forecasts in real time but only with very specific parameterizations; and (4) the conclusions we can

draw about bias in survey forecasts are heavily dependent on the choice of actuals for data that are subject to revisions. The main contributions of this paper to the literature on the rationality of forecasts are to provide more evidence about the sub-sample variation in estimates of bias and to provide a more-detailed examination of forecast-improvement exercises than has been done before, including the use of shrinkage.

Why might forecasters show periodic bouts of bias in their forecasts? As Farmer, Nakamura and Steinsson (2023) suggest, forecasters may not know the data-generating process at a given date but learn more about it over time. Our results are consistent with their theoretical model—forecasters do the best they can with a changing structure of the economy, and biases appear from time to time but disappear once forecasters understand the structural change.

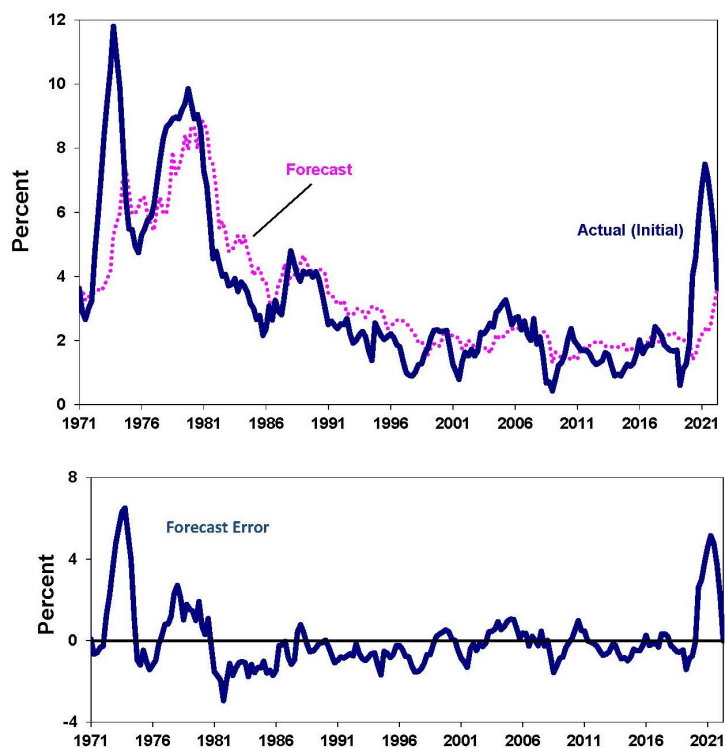
V. Tables and Graphs

FIGURE 1. MEAN ONE-YEAR-AHEAD OUTPUT GROWTH FORECAST AND ACTUAL



Note: The upper panel shows one-year-ahead output growth forecasts from the SPF (Forecast) and realized values based on the initial data release, labeled Actual (Initial). The bottom panel shows the forecast error, measured as actual minus forecast. The date shown on the horizontal axis is the forecast date. The sample period is based on forecasts made from 1971Q1 to 2022Q2. Note some large forecast errors and some persistent errors.

FIGURE 2. MEAN ONE-YEAR-AHEAD INFLATION FORECAST AND ACTUAL



Note: The upper panel shows one-year-ahead inflation forecasts from the SPF (Forecast) and realized values based on the initial data release, labeled Actual (Initial). The bottom panel shows the forecast error, measured as actual minus forecast. The date shown on the horizontal axis is the forecast date. The sample period is based on forecasts made from 1971Q1 to 2022Q2. Note some large forecast errors and some persistent errors.

TABLE 1—TEST FOR BIAS, ONE-YEAR AHEAD, BASED ON MEAN SPF FORECAST, FULL SAMPLE

Actual	Mean Error	p -value	Standard Error
Output growth			
Initial	−0.45	0.07	0.24
First revision	−0.42	0.09	0.25
First final	−0.41	0.09	0.25
Annual	−0.37	0.15	0.26
Pre-benchmark	−0.41	0.10	0.25
Last	−0.17	0.48	0.24
Inflation			
Initial	0.03	0.14	0.21
First revision	0.04	0.21	0.22
First final	0.05	0.25	0.22
Annual	0.05	0.24	0.20
Pre-benchmark	0.13	0.57	0.23
Last	0.06	0.27	0.22

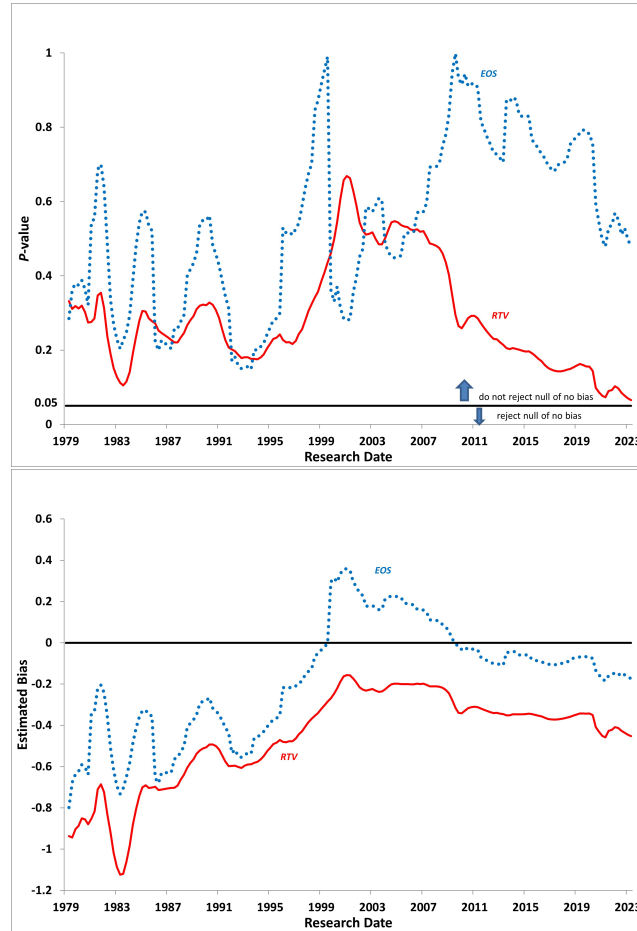
Note: The table shows the results of the zero-mean forecast-error test for output growth and inflation forecasts using the six different alternative measures of realized values. The sample uses SPF forecasts from 1971Q1 to 2022Q2. The p -value is a standard t -test for the null hypothesis that the mean forecast error is zero. Standard errors are adjusted following the Newey and West (1987) procedure.

TABLE 2—TEST FOR BIAS, ONE-YEAR AHEAD, BASED ON MEAN SPF FORECAST, PRE-COVID SAMPLE

Actual	Mean Error	p -value	Standard Error
Output growth			
Initial	−0.34	0.16	0.24
First revision	−0.31	0.20	0.24
First final	−0.31	0.21	0.24
Annual	−0.32	0.20	0.25
Pre-benchmark	−0.33	0.19	0.25
Last	−0.07	0.77	0.24
Inflation			
Initial	−0.11	0.59	0.20
First revision	−0.09	0.65	0.20
First final	−0.08	0.68	0.20
Annual	0.01	0.97	0.20
Pre-benchmark	−0.01	0.96	0.21
Last	−0.09	0.65	0.20

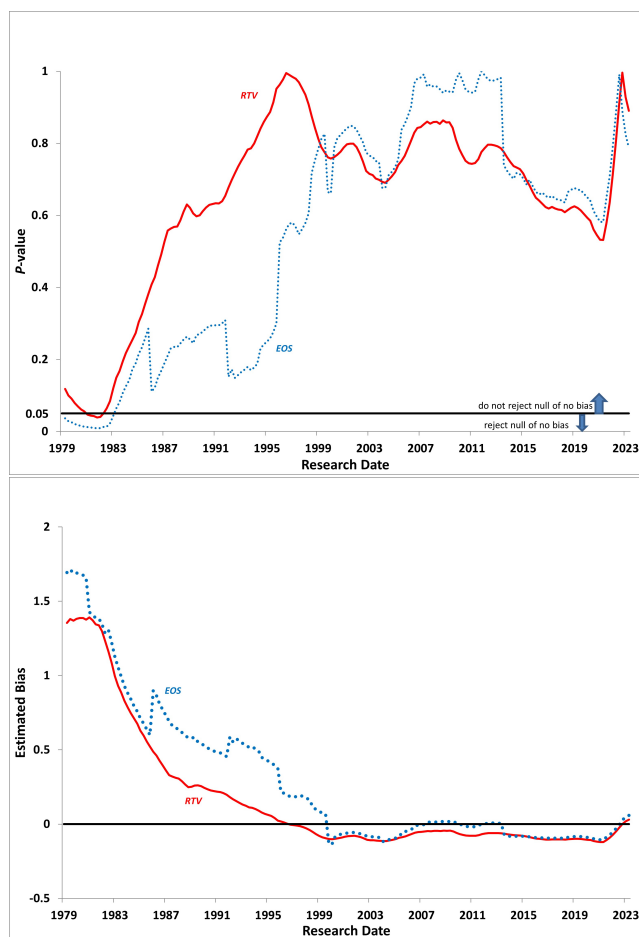
Note: The table shows the results of the zero-mean forecast-error test for output growth and inflation forecasts using the six different alternative measures of realized values. The sample is 1971Q1 to 2018Q4. The p -value is a standard t -test for the null hypothesis that the mean forecast error is zero. Standard errors are adjusted following the Newey and West (1987) procedure.

FIGURE 3. P -VALUES FOR BIAS AND ESTIMATED BIAS IN OUTPUT GROWTH FORECASTS FOR SAMPLE OBSERVED BY RESEARCHER AT ALTERNATIVE RESEARCH DATES

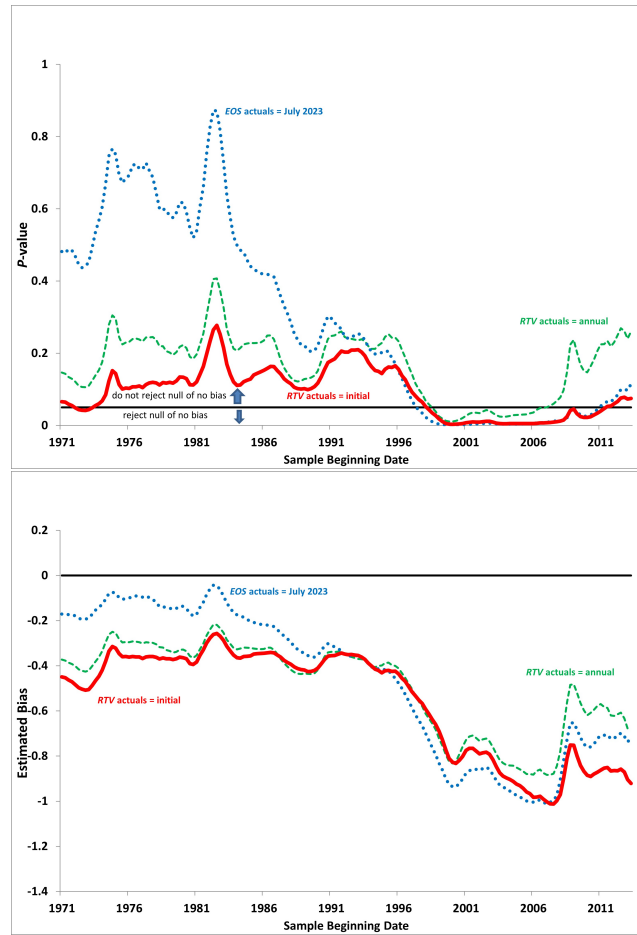


Note: The upper panel shows the p -values from the zero-mean forecast error test that would have been calculated by a researcher using data at the date shown on the horizontal axis. Each line corresponds to a different measure of actual output growth: using latest-available data at each research date, or using just initial releases at each research date. The horizontal line shows where $\rho = 0.05$. The lower panel shows the estimated bias in the forecast at each date. The research dates are 1979Q2 to 2023Q2.

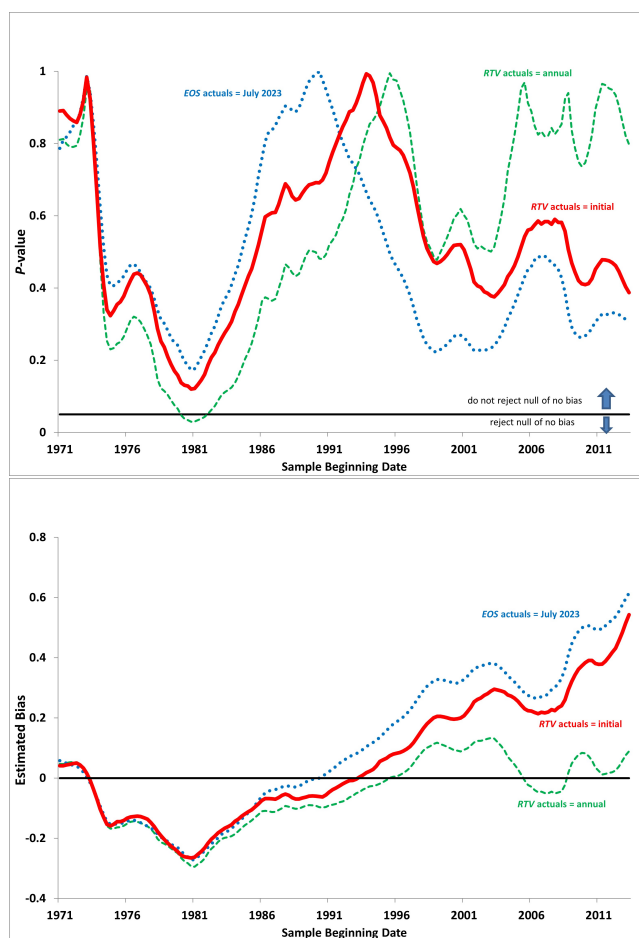
FIGURE 4. *P*-VALUES FOR BIAS AND ESTIMATED BIAS IN INFLATION FORECASTS FOR SAMPLE OBSERVED BY RESEARCHER AT ALTERNATIVE RESEARCH DATES



Note: The upper panel shows the *p*-values from the zero-mean forecast error test that would have been calculated by a researcher using data at the date shown on the horizontal axis. Each line corresponds to a different measure of actual output growth: using latest-available data at each research date, or using just initial releases at each research date. The horizontal line shows where $\rho = 0.05$. The lower panel shows the estimated bias in the forecast at each date. The research dates are 1979Q2 to 2023Q2.

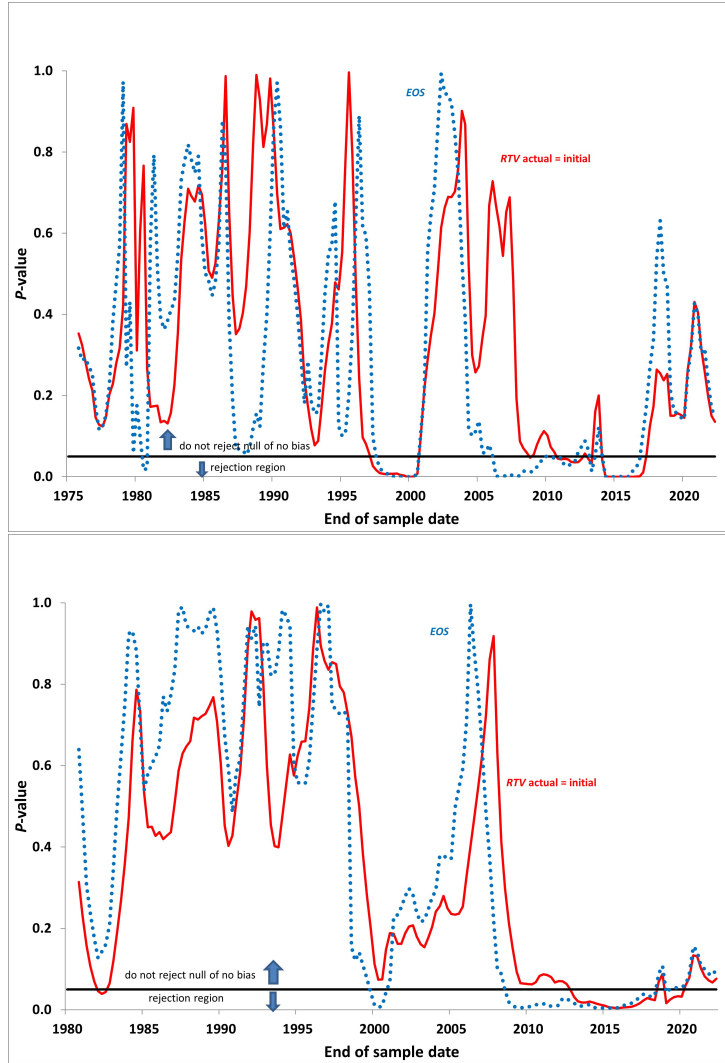
FIGURE 5. *P*-VALUES FOR BIAS IN OUTPUT FORECASTS FOR SAMPLE STARTING AT ALTERNATIVE DATES

Note: The upper panel shows the *p*-values from the zero-mean forecast error test that would have been calculated by a researcher using data through 2023Q2, with a sample starting at the date shown on the horizontal axis. Each line corresponds to a different method or a different measure of actual inflation: the *EOS* method using the data as of August 2023, and the *RTV* method using initial actuals or annual actuals. The lower panel shows the estimated bias in the forecast at each date. The sample beginning dates vary from 1971Q1 to 2013Q2.

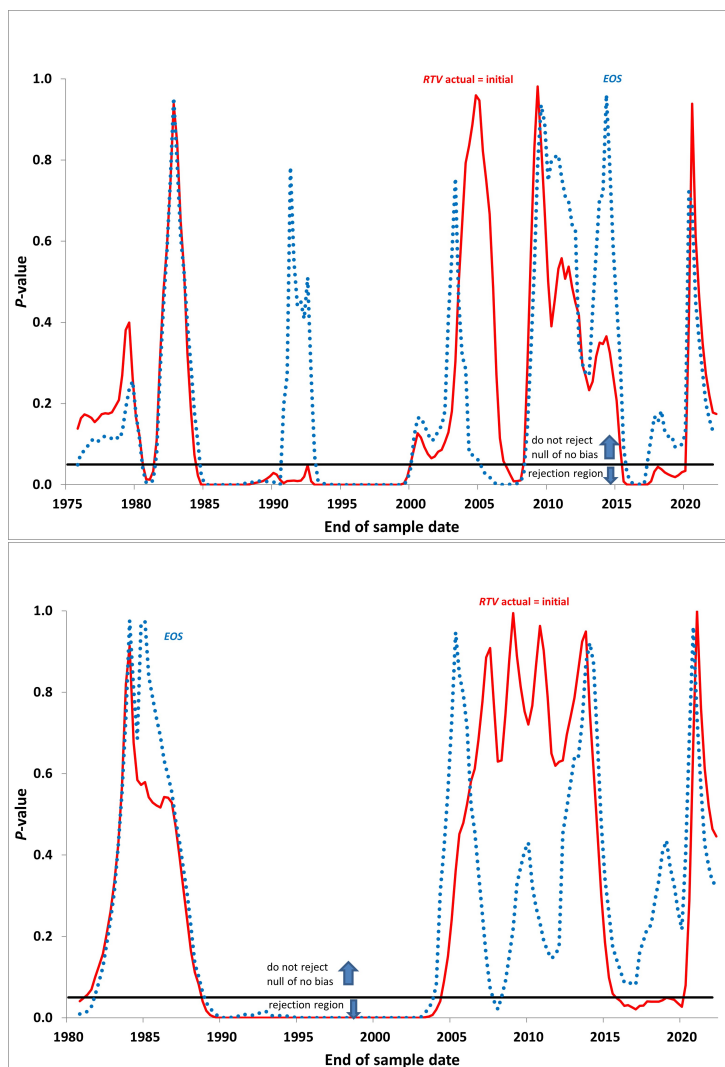
FIGURE 6. P -VALUES FOR BIAS IN INFLATION FORECASTS FOR SAMPLE STARTING AT ALTERNATIVE DATES

Note: The upper panel shows the p -values from the zero-mean forecast error test that would have been calculated by a researcher using data through 2023Q2, with a sample starting at the date shown on the horizontal axis. Each line corresponds to a different method or a different measure of actual inflation: the *EOS* method using the data as of August 2023, and the *RTV* method using initial actuals or annual actuals. The lower panel shows the estimated bias in the forecast at each date. The sample beginning dates vary from 1971Q1 to 2013Q2.

FIGURE 7. P -VALUES FOR BIAS IN FORECASTS FOR OUTPUT GROWTH IN ROLLING 5-YEAR AND 10-YEAR WINDOWS

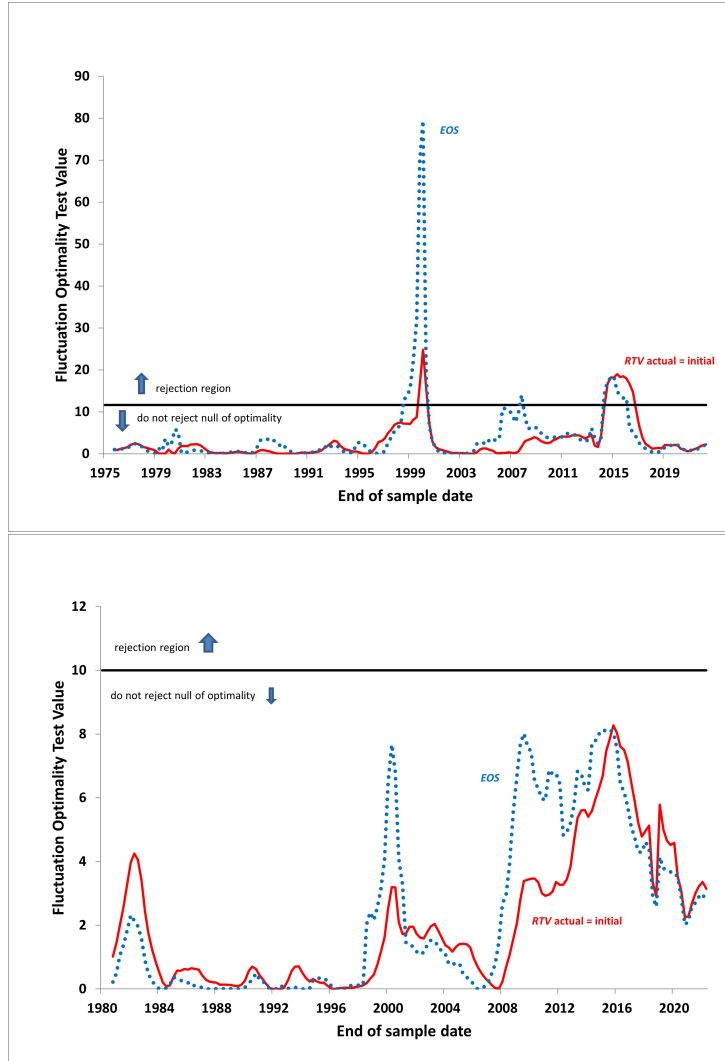


Note: The upper panel shows the p -values from the zero-mean forecast error test that would have been calculated by a researcher using rolling 5-year samples of data on output growth forecasts. Each line corresponds to a different method or a different measure of actual inflation: the *EOS* method or the *RTV* method using initial actuals. The lower panel shows the same concept for 10-year rolling windows. The sample ending forecast dates vary from 1975Q4 to 2022Q2.

FIGURE 8. *P*-VALUES FOR BIAS IN FORECASTS FOR INFLATION IN ROLLING 5-YEAR AND 10-YEAR WINDOWS

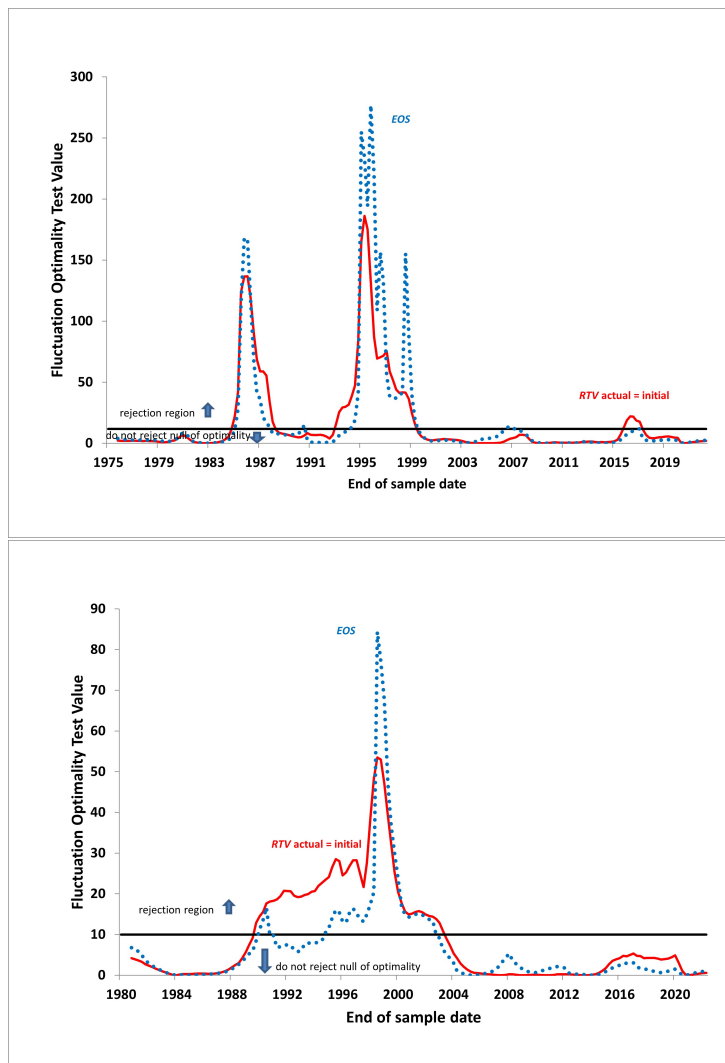
Note: The upper panel shows the *p*-values from the zero-mean forecast error test that would have been calculated by a researcher using rolling 5-year samples of data on inflation forecasts, while the bottom panel uses 10-year rolling windows. Each line corresponds to a different method or a different measure of actual inflation: the *EOS* method or the *RTV* method using initial actuals. The sample ending forecast dates vary from 1975Q4 to 2022Q2.

FIGURE 9. OUTPUT GROWTH: FLUCTUATION OPTIMALITY TEST WITH 5-YEAR AND 10-YEAR ROLLING WINDOWS



Note: The upper panel shows the values from the fluctuation optimality test with 5-year rolling windows, while the bottom panel uses 10-year rolling windows. Each line corresponds to a different method or a different measure of actual inflation: the *EOS* method or the *RTV* method using initial actuals. The sample ending forecast dates vary from 1975Q4 to 2022Q2 for 5-year rolling windows, and 1980Q4 to 2022Q2 for 10-year rolling windows.

FIGURE 10. INFLATION: FLUCTUATION OPTIMALITY TEST WITH 5-YEAR AND 10-YEAR ROLLING WINDOWS



Note: The upper panel shows the values from the fluctuation optimality test with 5-year rolling windows, while the bottom panel uses 10-year rolling windows. Each line corresponds to a different method or a different measure of actual inflation: the *EOS* method or the *RTV* method using initial actuals. The sample ending forecast dates vary from 1975Q4 to 2022Q2 for 5-year rolling windows, and 1980Q4 to 2022Q2 for 10-year rolling windows.

TABLE 3—*RRMSFEs* AND *P*-VALUES FOR FORECAST IMPROVEMENT EXERCISES, *RTV* METHOD WITH ACTUALS = INITIAL

Window Size:	5-year	10-year
Output Growth <i>RMSFE</i> = 1.45		
Adjust every period	1.10 [0.22]	1.11 [0.17]
Adjust when $\rho < 0.05$	0.98 [0.51]	1.01 [0.69]
Adjust when <i>FO</i> test rejects	1.00 [0.83]	NA
Inflation <i>RMSFE</i> = 0.79		
Adjust every period	1.07 [0.59]	1.18 [0.35]
Adjust when $\rho < 0.05$	0.94 [0.56]	0.93 [0.39]
Adjust when <i>FO</i> test rejects	0.92 [0.30]	0.91 [0.17]

Note: The table shows relative-root-mean-squared errors (*RRMSFE*) and *p*-values [in square brackets] for output growth and inflation forecasts in forecast-improvement exercises. The sample consists of one-year-ahead SPF forecasts made from 1982Q1 to 2018Q4.

TABLE 4—*RRMSFEs* AND *P*-VALUES FOR FORECAST IMPROVEMENT EXERCISES, *RTV* METHOD WITH ACTUALS = INTIAL, WITH SHRINKAGE

Window Size:	5-year	10-year
Output Growth <i>RMSFE</i> = 1.45		
Adjust every period	1.02 [0.56]	1.04 [0.34]
Adjust when $\rho < 0.05$	0.98 [0.23]	1.00 [0.77]
Adjust when <i>FO</i> test rejects	1.00 [0.89]	NA
Inflation <i>RMSFE</i> = 0.79		
Adjust every period	0.96 [0.58]	1.04 [0.69]
Adjust when $\rho < 0.05$	0.91 [0.05]	0.92 [0.09]
Adjust when <i>FO</i> test rejects	0.92 [0.06]	0.92 [0.05]

Note: The table shows relative-root-mean-squared errors (*RRMSFE*) and *p*-values [in square brackets] for output growth and inflation forecasts in forecast-improvement exercises with shrinkages (weight of one-half on bias coefficient). The sample consists of one-year-ahead SPF forecasts made from 1982Q1 to 2018Q4.

TABLE 5—*RRMSFEs* AND *P*-VALUES FOR FORECAST IMPROVEMENT EXERCISES, *EOS* METHOD WITH ACTUALS = INITIAL

Window Size:	5-year	10-year
Output Growth <i>RMSFE</i> = 1.45		
Adjust every period	1.16 [0.17]	1.10 [0.08]
Adjust when $\rho < 0.05$	0.98 [0.73]	1.04 [0.27]
Adjust when <i>FO</i> test rejects	0.98 [0.32]	NA
Inflation <i>RMSFE</i> = 0.79		
Adjust every period	1.13 [0.35]	1.23 [0.25]
Adjust when $\rho < 0.05$	1.03 [0.76]	0.94 [0.30]
Adjust when <i>FO</i> test rejects	0.94 [0.31]	0.98 [0.59]

Note: The table shows relative-root-mean-squared errors (*RRMSFE*) and *p*-values [in square brackets] for output growth and inflation forecasts in forecast-improvement exercises, using the *EOS* method with actuals = initial. The sample consists of one-year-ahead SPF forecasts made from 1982Q1 to 2018Q4.

TABLE 6—*RRMSFEs* AND *P*-VALUES FOR FORECAST IMPROVEMENT EXERCISES, *EOS* METHOD WITH ACTUALS = INITIAL AND SHRINKAGE

Window Size:	5-year	10-year
Output Growth <i>RMSFE</i> = 1.45		
Adjust every period	1.04 [0.42]	1.02 [0.42]
Adjust when $\rho < 0.05$	0.97 [0.23]	1.00 [0.96]
Adjust when <i>FO</i> test rejects	0.98 [0.22]	NA
Inflation <i>RMSFE</i> = 0.79		
Adjust every period	1.00 [0.94]	1.07 [0.44]
Adjust when $\rho < 0.05$	0.95 [0.27]	0.94 [0.08]
Adjust when <i>FO</i> test rejects	0.94 [0.05]	0.97 [0.23]

Note: The table shows relative-root-mean-squared errors (*RRMSFE*) and *p*-values [in square brackets] for output growth and inflation forecasts in forecast-improvement exercises, using the *EOS* method with actuals = initial and shrinkage of the coefficient in the bias regression by one half. The sample consists of one-year-ahead SPF forecasts made from 1982Q1 to 2018Q4.

TABLE 7—*RRMSFEs* AND *P*-VALUES FOR FORECAST IMPROVEMENT EXERCISES, *EOS* METHOD WITH ACTUALS = PRE-BENCHMARK

Window Size:	5-year	10-year
Output Growth <i>RMSFE</i> = 1.57		
Adjust every period	1.18 [0.07]	1.12 [0.02]
Adjust when $\rho < 0.05$	1.01 [0.80]	1.04 [0.23]
Adjust when <i>FO</i> test rejects	1.00 [0.86]	NA
Inflation <i>RMSFE</i> = 0.91		
Adjust every period	1.13 [0.29]	1.22 [0.20]
Adjust when $\rho < 0.05$	1.02 [0.84]	0.97 [0.59]
Adjust when <i>FO</i> test rejects	0.96 [0.35]	0.98 [0.55]

Note: The table shows relative-root-mean-squared errors (*RRMSFE*) and *p*-values [in square brackets] for output growth and inflation forecasts in forecast-improvement exercises, using the *EOS* method with actuals = pre-benchmark. The sample consists of one-year-ahead SPF forecasts made from 1982Q1 to 2018Q4.

TABLE 8—*RRMSFEs* AND *P*-VALUES FOR FORECAST IMPROVEMENT EXERCISES, *EOS* METHOD WITH ACTUALS = PRE-BENCHMARK AND SHRINKAGE

Window Size:	5-year	10-year
Output Growth <i>RMSFE</i> = 1.57		
Adjust every period	1.06 [0.24]	1.04 [0.18]
Adjust when $\rho < 0.05$	0.99 [0.68]	1.00 [0.85]
Adjust when <i>FO</i> test rejects	1.00 [0.87]	NA
Inflation <i>RMSFE</i> = 0.91		
Adjust every period	1.01 [0.88]	1.08 [0.36]
Adjust when $\rho < 0.05$	0.96 [0.18]	0.96 [0.21]
Adjust when <i>FO</i> test rejects	0.95 [0.03]	0.98 [0.55]

Note: The table shows relative-root-mean-squared errors (*RRMSFE*) and *p*-values [in square brackets] for output growth and inflation forecasts in forecast-improvement exercises, using the *EOS* method with actuals = pre-benchmark and shrinkage of the coefficient in the bias regression by one half. The sample consists of one-year-ahead SPF forecasts made from 1982Q1 to 2018Q4.

REFERENCES

- Aiolfi, Marco, Carlos Capistran, and Allan Timmermann.** 2011. “Forecast Combinations.” In *The Oxford Handbook of Economic Forecasting*, ed. Michael P. Clements and David F. Hendry, Chapter 12, 355–388. Oxford University Press.
- Ang, Andrew, Geert Bekaert, and Min Wei.** 2007. “Do Macro Variables, Asset Markets, or Surveys Forecast Inflation Better?” *Journal of Monetary Economics*, 54: 1163–1212.
- Bonham, Carl, and Douglas C. Dacy.** 1991. “In Search of a Strictly Rational Forecast.” *Review of Economics and Statistics*, 73: 245–253.
- Carroll, Christopher D.** 2003. “Macroeconomic Expectations of Households and Professional Forecasters.” *Quarterly Journal of Economics*, 118: 269–298.
- Clark, Todd E., and Michael W. McCracken.** 2009. “Tests of Equal Predictive Ability with Real-Time Data.” *Journal of Business and Economic Statistics*, 27: 441–454.
- Croushore, Dean.** 2010. “An Evaluation of Inflation Forecasts from Surveys using Real-Time Data.” *B.E. Journal of Macroeconomics*, 10(1): article 10.
- Croushore, Dean.** 2011. “Frontiers of Real-Time Data Analysis.” *Journal of Economic Literature*, 49(1): 72–100. Federal Reserve Bank of Philadelphia working paper No. 08-4.
- Croushore, Dean, and Tom Stark.** 2001. “A Real-Time Data Set for Macroeconomists.” *Journal of Econometrics*, 105: 111–130.
- Croushore, Dean, and Tom Stark.** 2019. “Fifty Years of the Survey of Professional Forecasters.” *Federal Reserve Bank of Philadelphia Economic Insights*, 1–11.

- Diebold, Francis X., and Roberto S. Mariano.** 1995. “Comparing Predictive Accuracy.” *Journal of Business and Economic Statistics*, 13: 253–263.
- Elliott, Graham, Ivana Komunjer, and Allan Timmermann.** 2008. “Biases in Macroeconomic Forecasts: Irrationality or Asymmetric Loss?” *Journal of the European Economic Association*, 6: 122–157.
- Eva, Kenneth, and Fabian Winkler.** 2023. “A Comprehensive Empirical Evaluation of Biases in Expectation Formation.” Working Paper, Federal Reserve Board.
- Farmer, Leland E., Emi Nakamura, and Jon Steinsson.** 2023. “Learning About the Long Run.” Working Paper, National Bureau of Economic Research.
- Giacomini, Raffaella, and Barbara Rossi.** 2010. “Forecast Comparisons in Unstable Environments.” *Journal of Applied Econometrics*, 25: 595–620.
- Hafer, R.W., and Scott E. Hein.** 1985. “On the Accuracy of Time-Series, Interest Rate, and Survey Forecasts of Inflation.” *Journal of Business*, 58: 377–398.
- Keane, Michael P., and David E. Runkle.** 1990. “Testing the Rationality of Price Forecasts: New Evidence From Panel Data.” *American Economic Review*, 80: 714–735.
- Kishor, N. Kundan, and Evan F. Koenig.** 2012. “VAR Estimation and Forecasting When Data Are Subject to Revision.” *Journal of Business and Economic Statistics*, 30: 181–190. Federal Reserve Bank of Dallas Economic Research Working Paper No. 0501.
- Koenig, Evan, Sheila Dolmas, and Jeremy Piger.** 2003. “The Use and Abuse of ‘Real-Time’ Data in Economic Forecasting.” *Review of Economics and Statistics*, 85: 618–628.

- Mankiw, N. Gregory, and Matthew D. Shapiro.** 1986. "Do We Reject Too Often? Small Sample Bias in Tests of Rational Expectations Models." *Economics Letters*, 20: 139–145.
- Mankiw, N. Gregory, Ricardo Reis, and Justin Wolfers.** 2003. "Disagreement About Inflation Expectations." *NBER Macroeconomics Annual*, 209–248.
- Mincer, Jacob A., and Victor Zarnowitz.** 1969. "The Evaluation of Economic Forecasts." In *Economic Forecasts and Expectations.* , ed. Jacob Mincer. New York:National Bureau of Economic Research.
- Newey, Whitney K., and Kenneth D. West.** 1987. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica*, 55: 703–708.
- Romer, Christina D., and David H. Romer.** 2000. "Federal Reserve Information and the Behavior of Interest Rates." *American Economic Review*, 90(3): 429–457.
- Rossi, Barbara.** 2006. "Are Exchange Rates Really Random Walks? Some Evidence Robust to Parameter Instability." *Macroeconomic Dynamics*, 10: 20–38.
- Rossi, Barbara, and Tatevik Sekhposyan.** 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. jae.2440.
- Rudebusch, Glenn D., and John C. Williams.** 2009. "Forecasting Recessions: The Puzzle of the Enduring Power of the Yield Curve." *Journal of Business and Economic Statistics*, 27(4): 492–503.
- Stock, James H., and Mark W. Watson.** 2003. "Forecasting Output and Inflation: The Role of Asset Prices." *Journal of Economic Literature*, 41: 788–829.

- Su, Vincent, and Josephine Su.** 1975. "An Evaluation of ASA/NBER Business Outlook Survey Forecasts." *Explorations in Economic Research*, 2: 588–618.
- Zarnowitz, Victor.** 1985. "Rational Expectations and Macroeconomic Forecasts." *Journal of Business and Economic Statistics*, 3: 293–311.

APPENDIX

This Appendix contains the dates of the annual and pre-benchmark vintages, as well as some extra tables that might be interesting but would make the paper too long.

Annual Revision Dates for Quarterly National Accounts

The first annual revision occurred at the end of July each year (and are thus in the August monthly vintage of the RTDSM) pertaining to the prior calendar year every year except for the following. For example, the first annual revision of the data for 1965 was at the end of July 1966 and recorded in the 1966M8 vintage of the RTDSM. The exceptions are given in this table:

TABLE A1—ANNUAL REVISION DATES FOR QUARTERLY NATIONAL ACCOUNTS

Year Revised	First Annual Revision Date in RTDSM
1974	1976M2
1979	1981M1
1980	1981M8
1984	1986M1
1990	1991M12
1994	1996M1
1998	1999M11
2002	2003M12

Benchmark Revision Dates for Quarterly National Accounts

TABLE A2—PRE-BENCHMARK-REVISION RTDSM MONTHLY DATES

Benchmark Revision Date	Pre-Benchmark Date	Last Observation
1976M2	1976M1	1975Q3
1981M1	1980M12	1980Q3
1986M1	1985M12	1985Q3
1991M12	1991M11	1991Q3
1996M1	1995M12	1995Q3
1999M11	1999M10	1999Q2
2003M12	2003M11	2003Q3
2009M8	2009M7	2009Q1
2013M8	2013M7	2013Q1
2018M8	2018M7	2018Q1

Table A3 shows summary forecast error statistics for various outcome measures, based on SPF forecasts from 1971Q1 to 2022Q2. Following that, Table A4 is a table with the same information but ending before COVID, with the 2018Q4 SPF survey as last forecast being evaluated.

TABLE A3—SUMMARY STATISTICS, MEAN ACROSS FORECASTERS, ONE-YEAR-AHEAD FORECASTS

	Output			Inflation		
Actual concept	<i>ME</i>	<i>RMSFE</i>	<i>MAE</i>	<i>ME</i>	<i>RMSFE</i>	<i>MAE</i>
Initial	−0.45	2.16	1.44	0.03	1.52	1.01
First revision	−0.42	2.16	1.44	0.04	1.55	1.03
First final	−0.41	2.15	1.44	0.05	1.54	1.01
Annual	−0.37	2.20	1.48	0.05	1.40	0.93
Pre-benchmark	−0.41	2.19	1.50	0.13	1.61	1.08
Last vintage	−0.17	2.07	1.49	0.06	1.51	1.07

Note: The table shows mean errors (*ME*), root-mean-squared forecast errors (*RMSFE*), and mean absolute errors (*MAE*) for output growth and inflation forecasts using six different alternative measures of realized values: the initial release, the first revision, the first-final (second revision) release, the annual release, the pre-benchmark release and the last vintage. The sample consists of SPF forecasts made from 1971Q1 to 2022Q2.

TABLE A4—SUMMARY STATISTICS, ONE-YEAR-AHEAD FORECASTS, MEAN ACROSS FORECASTERS, PRE-COVID

	Output			Inflation		
Actual concept	<i>ME</i>	<i>RMSFE</i>	<i>MAE</i>	<i>ME</i>	<i>RMSFE</i>	<i>MAE</i>
Initial	−0.34	1.93	1.31	−0.11	1.36	0.91
First revision	−0.31	1.94	1.31	−0.09	1.38	0.92
First final	−0.31	1.93	1.32	−0.08	1.37	0.91
Annual	−0.32	1.99	1.37	0.01	1.36	0.89
Pre-benchmark	−0.33	2.02	1.40	−0.01	1.44	0.98
Last vintage	−0.07	1.88	1.39	−0.09	1.32	0.96

Note: The table shows mean errors (*ME*), root-mean-squared forecast errors (*RMSFE*), and mean absolute errors (*MAE*) for output growth and inflation forecasts using six different alternative measures of realized values: the initial release, the first revision, the first-final (second revision) release, the annual release, the pre-benchmark release and the last vintage. The sample consists of SPF forecasts made from 1971Q1 to 2018Q4.