WORKING PAPERS
RESEARCH DEPARTMENT

WORKING PAPER NO. 12-9
FORECAST BIAS IN TWO DIMENSIONS

Dean Croushore

Professor of Economics and Rigsby Fellow
University of Richmond
and
Visiting Scholar
Federal Reserve Bank of Philadelphia

April 2012

RESEARCH DEPARTMENT, FEDERAL RESERVE BANK OF PHILADELPHIA

Ten Independence Mall, Philadelphia, PA 19106-1574 • www.philadelphiafed.org/research-and-data/

FORECAST BIAS IN TWO DIMENSIONS

Dean Croushore

Professor of Economics and Rigsby Fellow
University of Richmond

Visiting Scholar
Federal Reserve Bank of Philadelphia

April 2012

Thanks for useful comments and suggestions to participants at the Joint Statistical Meetings of the American Statistical Association, the Computational Economics and Finance meeting, the Society for Nonlinear Dynamics and Econometrics conference, the Workshop on Real-Time Data Analysis, the Southern Economic Association meetings, West Virginia University, and the Centre Interuniversitaire de Recherche en Economie Quantitative. This paper was written in part while the author was a visiting scholar at the Federal Reserve Bank of Philadelphia. The views expressed in this paper are those of the author and do not necessarily represent the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System. This paper is available free of charge at www.philadelphiafed.org/research-and-data/publications/working-papers/.

Please send comments to the author at Robins School of Business, 1 Gateway Road, University of Richmond, VA 23173, or e-mail: dcrousho@richmond.edu.

FORECAST BIAS IN TWO DIMENSIONS

ABSTRACT

Economists have tried to uncover stylized facts about people's expectations, testing whether such expectations are rational. Tests in the early 1980s suggested that expectations were biased, and some economists took irrational expectations as a stylized fact. But, over time, the results of tests that led to such a conclusion were reversed. In this paper, we examine how tests for bias in expectations, measured using the Survey of Professional Forecasters, have changed over time. In addition, key macroeconomic variables that are the subject of forecasts are revised over time, causing problems in determining how to measure the accuracy of forecasts. 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. Thus, our analysis takes place in two dimensions: across subsamples and with alternative measures of realized values of variables.

FORECAST BIAS IN TWO DIMENSIONS

INTRODUCTION

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.

Both Croushore (2010) and Giacomini and Rossi (2010) find substantial instability across subsamples in evaluations of forecasts. No global stylized facts appear to be available. 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.

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 and Stock-Watson. 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.

DATA

In this paper, we study two different variables: the growth rate of real output and the inflation rate as measured by the GDP price index. These are the two most studied economic variables, yet the stability of forecasts of these variables has not been studied before, except by Croushore (2010) for the inflation rate. 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 by the U.S. Bureau of Economic Analysis for the first time on a regular basis.

To study the ability of forecasters to provide accurate forecasts, we use the Survey of Professional Forecasters, SPF (see Croushore (1993)), which records the forecasts of a large number of private-sector forecasters. The literature studying the SPF forecasts has found that the SPF forecasts outperform macroeconomic models, even fairly sophisticated ones, as shown by Ang et al. (2007). The SPF has also been found to influence household expectations, as shown by Carroll (2003).

The SPF contains a number of different forecasts of output growth and inflation. For this paper, we choose to analyze the one-year-ahead forecasts, measured by the median forecast of the forecasters in the survey. While some arguments can be made that testing rational expectations is best done by examining the forecasts of individual forecasters (see Keane and Runkle (1990)), a more compelling argument is that the most accurate forecasts are provided by taking the median across the forecasters, as illustrated by Aiolfi et al. (2010). 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 median forecasts of output and inflation are reported in the SPF beginning with the fourth quarter of 1968. 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 study 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 four quarters. This variable is subject to less noise and presumably more economic causes than would be the case for studying the forecasts for a particular quarterly horizon. Of course, it is possible to combine information across horizons, as is done recently by Patton and Timmermann (2011), but an analysis across horizons would introduce a third dimension to our analysis, which is already complicated enough, so we leave this idea to future research.

We begin by looking at the forecasts and forecast errors in Figure 1a for output growth and Figure 1b for inflation.

Figure 1a
Median Output Growth Forecast from SPF
Using Initial Release as Actual

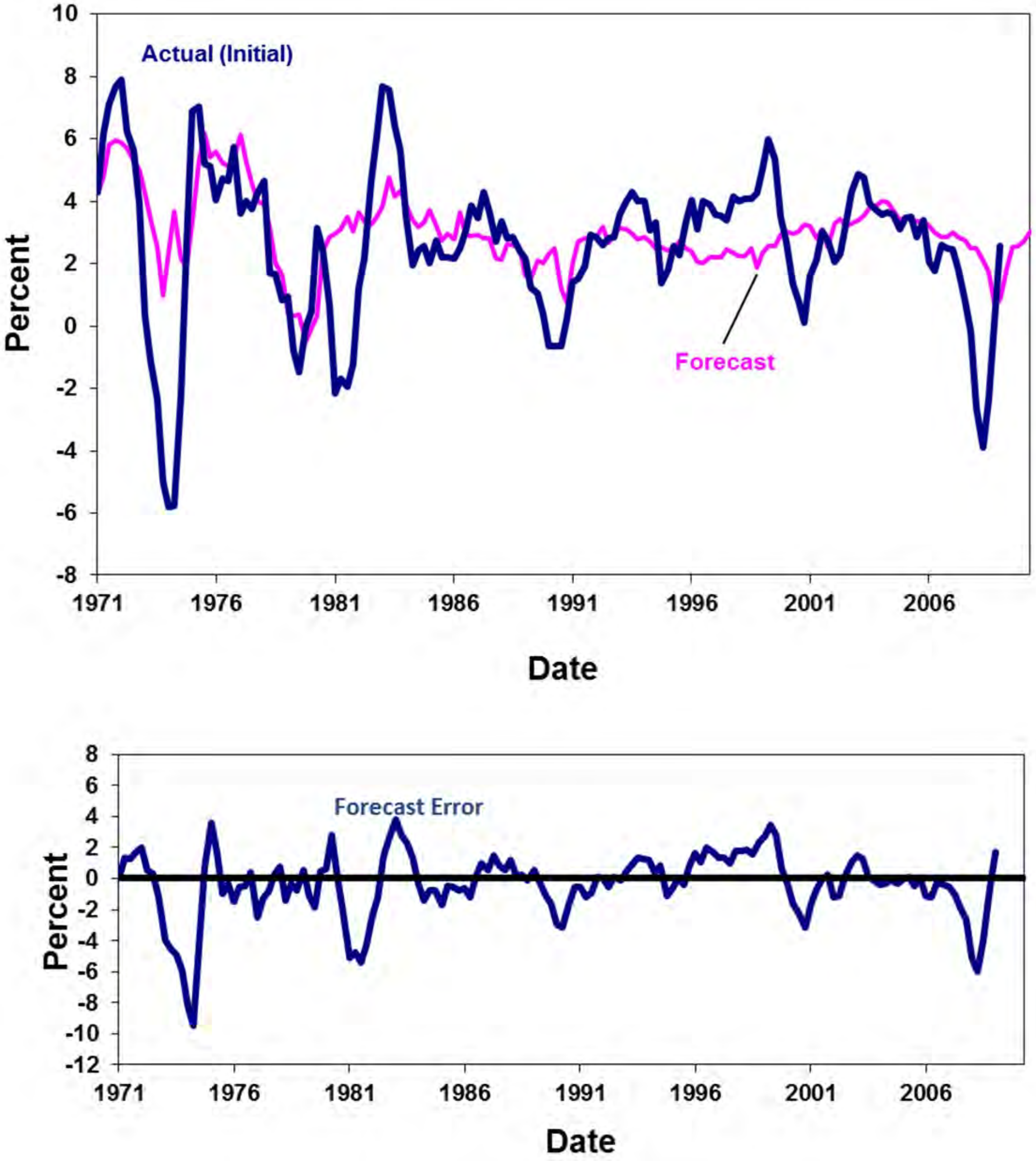
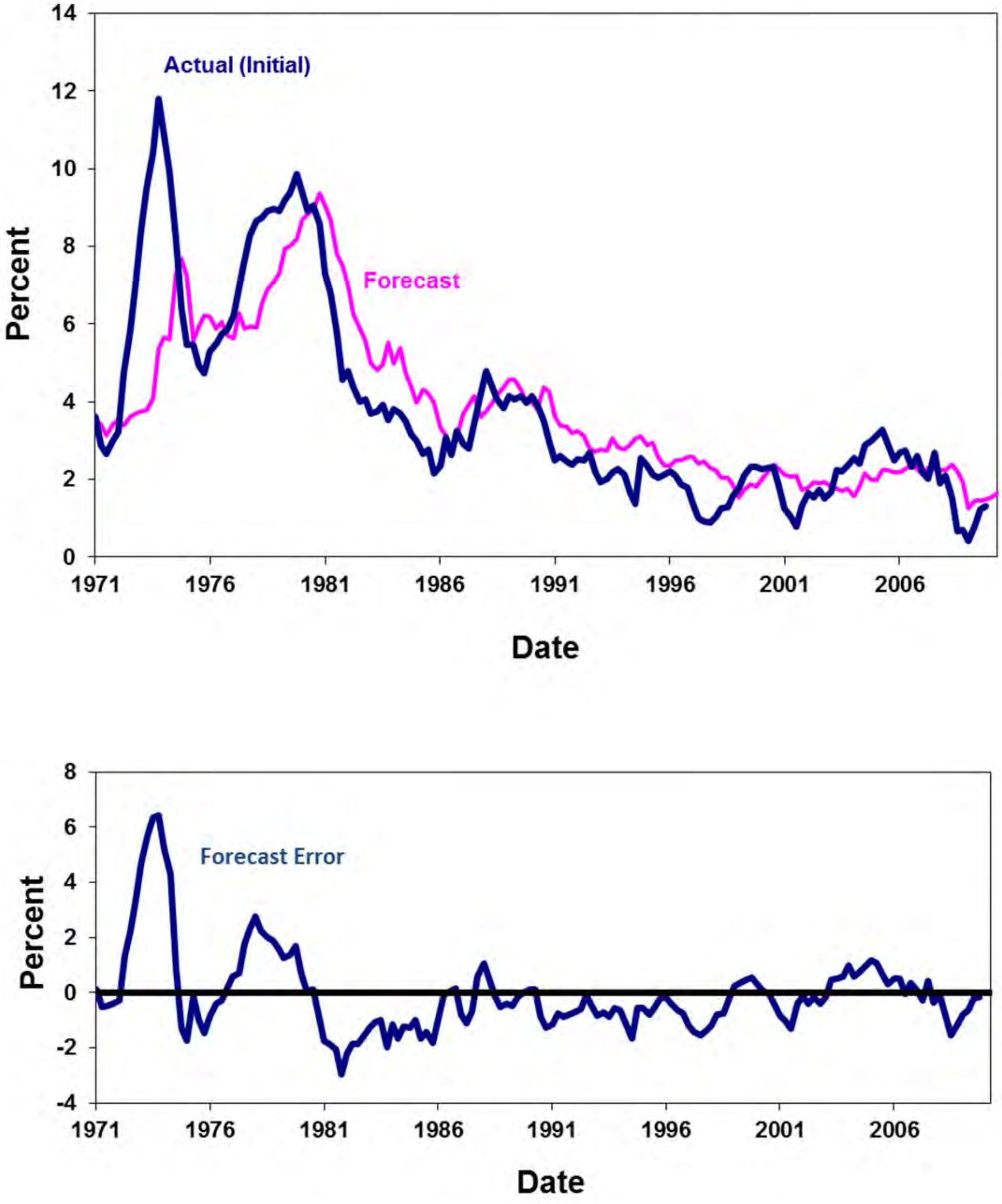


Figure 1b
Median Inflation Forecast from SPF
Using Initial Release as Actual



The figures are based on using the initial data release as actual; of course, other concepts of actual could be used. 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 Hansen and Hodrick (1980) and 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. Based on a review of the revision process, we will examine four different concepts for actual output and inflation: (1) the initial release; (2) the annual release, which is usually produced each year at the end of July; (3) 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 (4) the July 2010 version of the data, the latest-available vintage when the empirical work on this paper was begun. In years in which a benchmark release occurs, such 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.

To illustrate the size of the revisions, Figure 2a shows the revisions from the initial release to the annual release of the quarterly growth rate (shown at an annualized rate) of output and the revisions to the growth rate over the past four quarters; Figure 2b does the same for the inflation rate. The four-quarter growth rate (inflation rate) is shown for two reasons: (1) it is the main object of our study; and (2) it illustrates that large quarterly revisions do not entirely wash out over the four quarters. Revisions to quarterly output growth are as large as +6 percent and as small as –3 percent. Revisions to the growth rate of output over four quarters are around plus or minus two percent. For inflation, revisions to quarterly data are as large as +3 percent and as small as –2 percent; revisions to four-quarter inflation rates bounce between +1 percent and –1 percent.

In addition to the significant size of the revisions that is apparent in Figures 2a and 2b, revisions are nontrivial in several other aspects. Revisions to particular observations can be very large and persistent. In addition, long-term growth rates of macroeconomic variables, such as growth rates over five-year periods, can also be revised substantially. For more details on these revisions, see Croushore (2011).

Figure 2a
Revisions to GDP Growth Rate
Initial to Annual

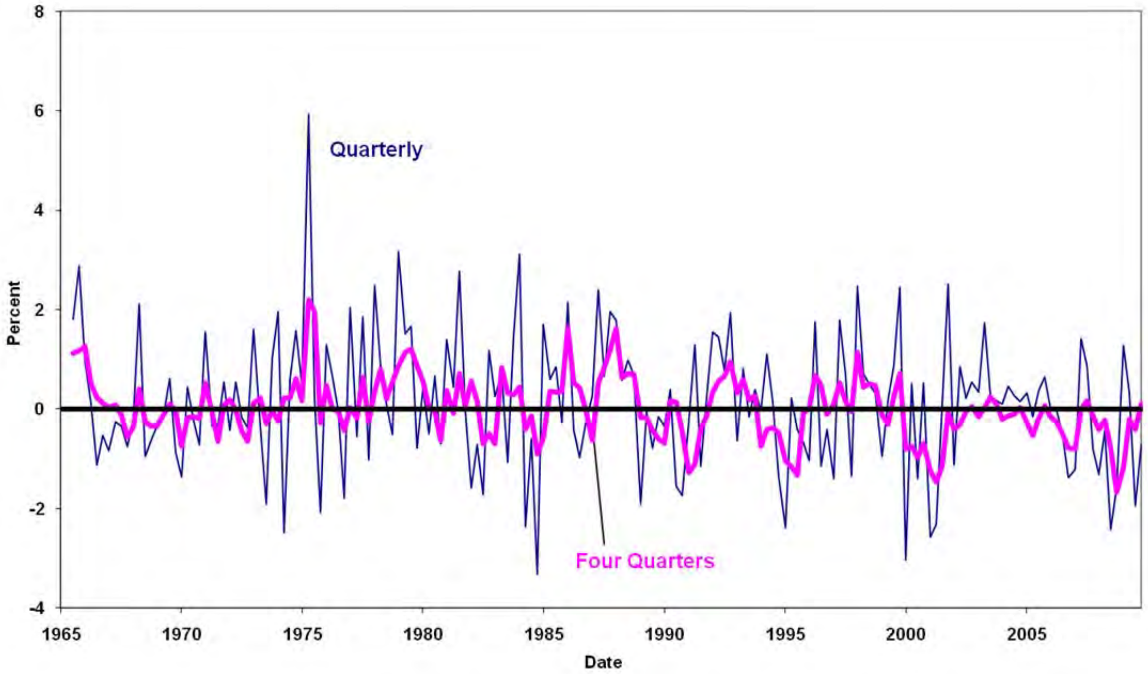
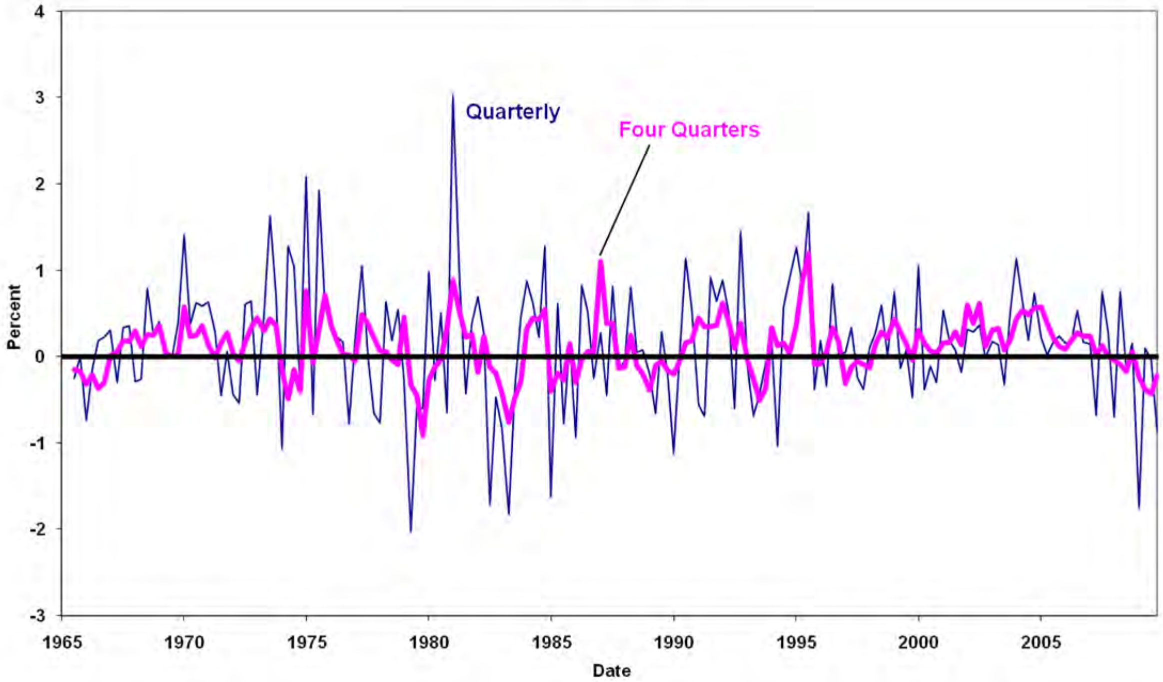


Figure 2b
Revisions to Inflation Rate
Initial to Annual



RESULTS

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-Zarnowitz (1969) test, which runs the regression:

$$A_t = \alpha + \beta F_t + \epsilon_t, \qquad (1)$$

where A_t is the actual value at time *t* and F_t is the forecast value. From the results of this regression, the researcher can test for the joint null hypothesis that α = 0 and β = 1. If that null hypothesis is rejected, the forecasts are thought to be biased.[1]

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. Defining $\varepsilon_t = A_t - F_t$, we run the regression:

$$\varepsilon_t = \alpha + \epsilon_t. \qquad (2)$$

In this equation, we are implicitly imposing that β = 1 in equation (1) and examining the properties of the forecast error itself, rather than studying whether the actual value of the variable increases one-for-one with the forecast.[2]

[1] We 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 et al. (2008), allow for the possibility that the loss function of the forecasters may be asymmetric, which implies that bias in forecasts may be optimal.

[2] We have also run the Mincer-Zarnowitz test to see how it performs, and the increased parameter uncertainty and small-sample bias lead to much worse results for the forecast-improvement exercises that we show later. So, in this paper we will show only the results for the zero-mean test.

We run the zero-mean test for both output growth and inflation, using all four versions of actuals: initial release, annual revision, pre-benchmark revision, and latest available. The results of this exercise are shown in Table 1. In each case, we show the mean forecast error $\hat{\alpha}$, the *p*-value testing whether the mean forecast error is significantly different from zero, and whether the null hypothesis of zero-mean forecast error is rejected or not.

Table 1: Test for Bias

Actual	**Mean Error**	***p*-value**	**Reject null?**
Output Growth			
Initial	-0.43	0.16	no
Annual	-0.43	0.17	no
Pre-benchmark	-0.41	0.20	no
July 2010	-0.14	0.64	no
Inflation			
Initial	-0.06	0.81	no
Annual	0.05	0.83	no
Pre-benchmark	0.03	0.92	no
July 2010	0.00	0.99	no

Table 1 shows that for all versions of actuals and for both variables, we never reject the null hypothesis of zero-mean forecast error. However, Croushore (2010) shows that results like this 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 1978, with data and forecasts made from 1971:Q1 to 1976:Q4. What conclusion about bias would she have drawn? We can ask the same question for a researcher standing at any date between 1978:Q2 and 2008:Q4.

To analyze these results, we consider the various measures of actuals that we used earlier, as well as another version: latest-available data. The latest-available-vintage actual is identical to the data that someone in real time would have downloaded from a database at the time. So, the latest-available vintage that we used in this paper is the July 2010 data. But a researcher standing in the second quarter of 1978 would have had a very different version of the latest-available data than the July 2010 vintage. So, we can collect a sequence of latest-available data sets at each date and call the actuals created using those data "latest-available" actuals.

The results of this exercise are shown in Figures 3a (for output) and 3b (for inflation). There are not many significant rejections of the null of no bias—only for a few cases when the sample ends 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. 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.

These general conclusions hold no matter which version of actual is used. However, for output growth, there are no rejections except for the initial actuals, so the choice of actuals matters. And for inflation, the choice of actuals has a significant influence on the sample periods under which the null of no bias is rejected. The sample periods when actual = initial, for which we reject unbiasedness, are much less than for other actuals, while there are many more sample periods for actuals = latest available, for which we reject unbiasedness.

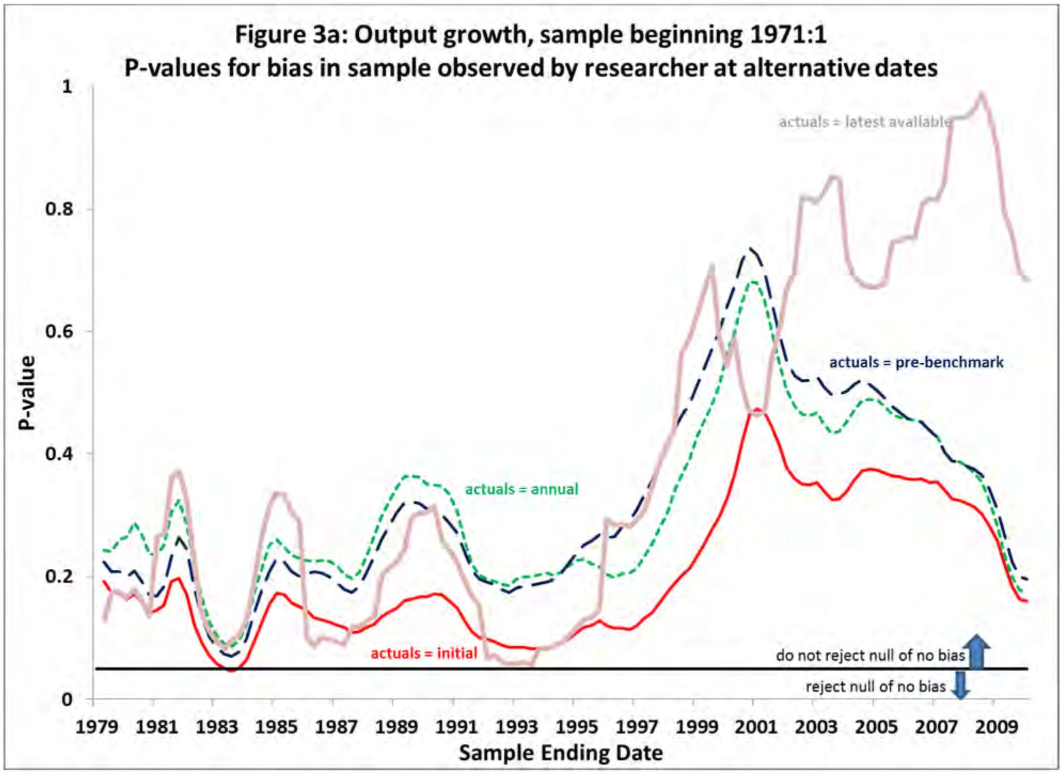


Figure 3a: Output growth, sample beginning 1971:1
P-values for bias in sample observed by researcher at alternative dates

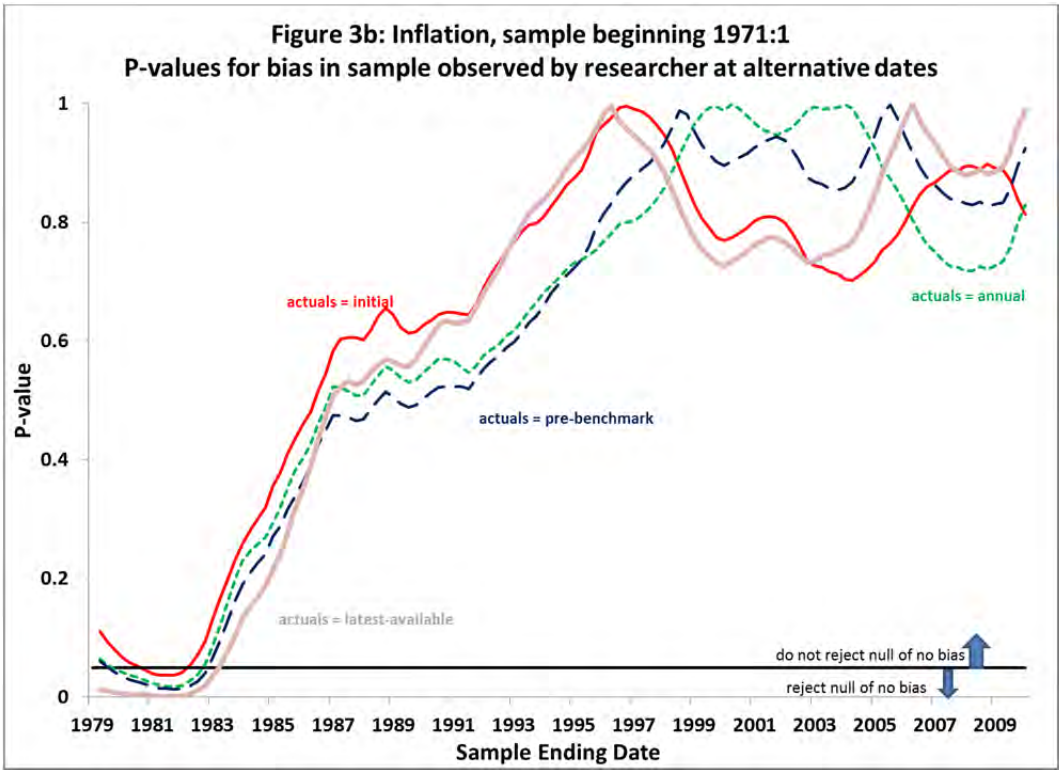
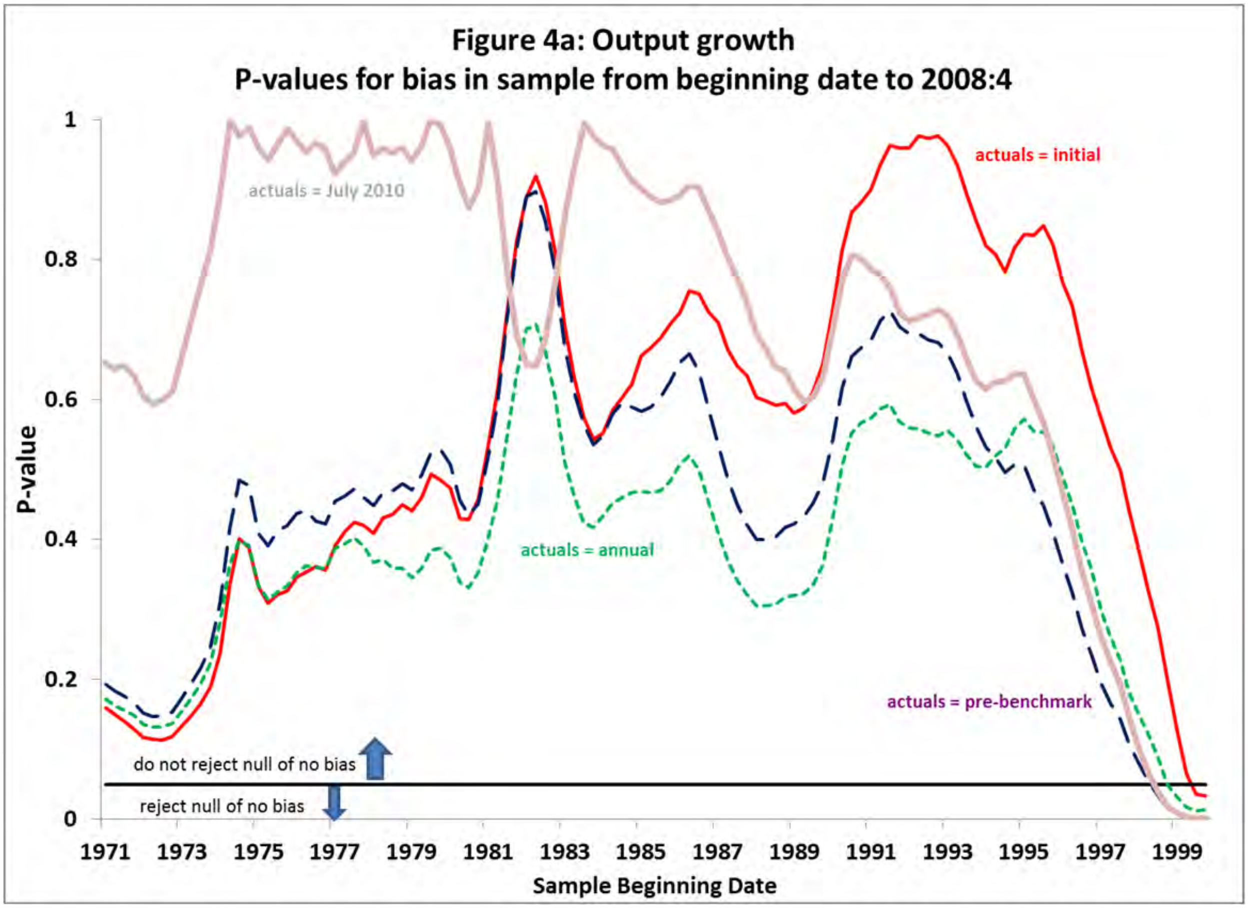
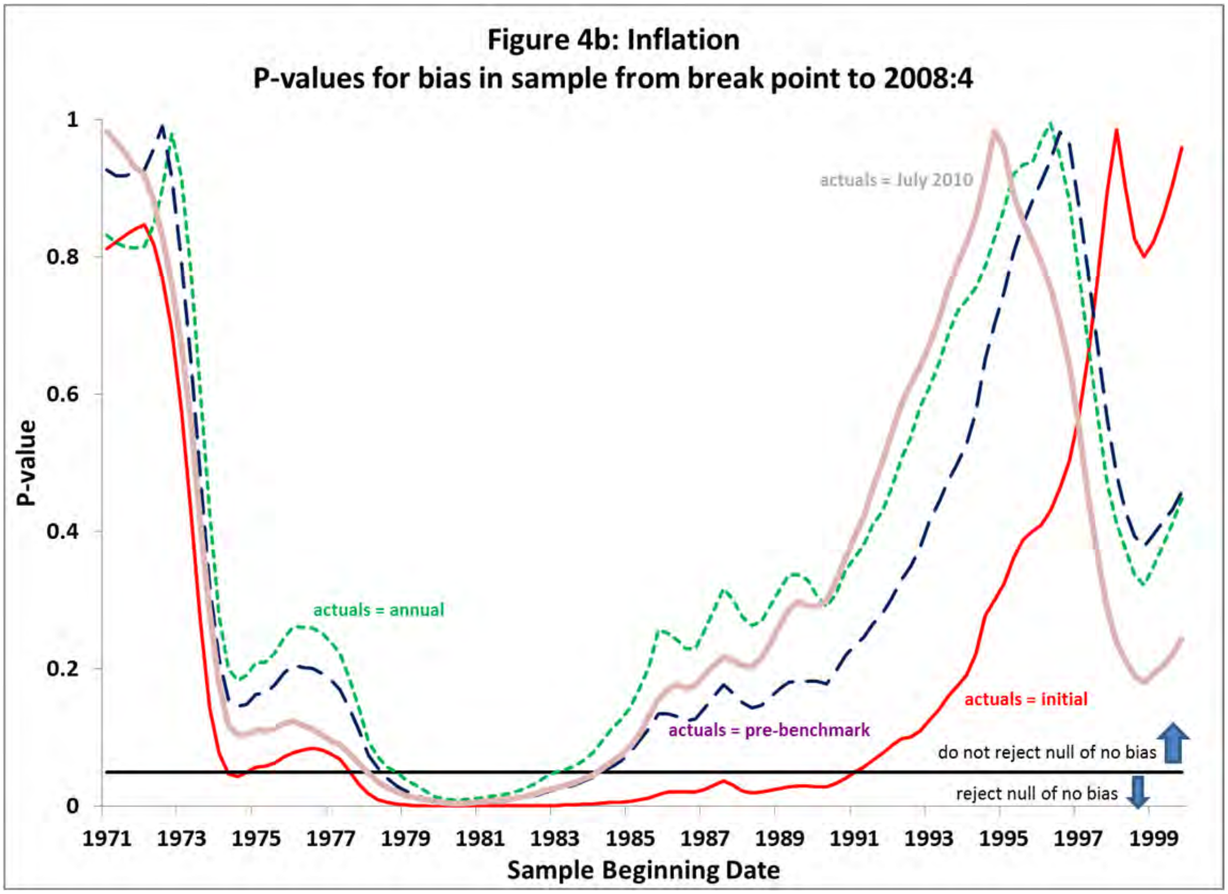


Figure 3b: Inflation, sample beginning 1971:1
P-values for bias in sample observed by researcher at alternative dates

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? So, consider subsamples that end in 2008:Q4, but begin at various dates after 1971:Q1. Figures 4a and 4b show the results of this exercise.

In these figures, we can see that we reject the null of no bias in many additional subsamples, especially for inflation. 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 initial actuals showing bias for many more subsamples than the other actuals.





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 test as before, but for rolling 5-year and 10-year windows. The p-values for this exercise are shown in Figures 5a and 5b (5-year windows) and 6a and 6b (10-year windows).

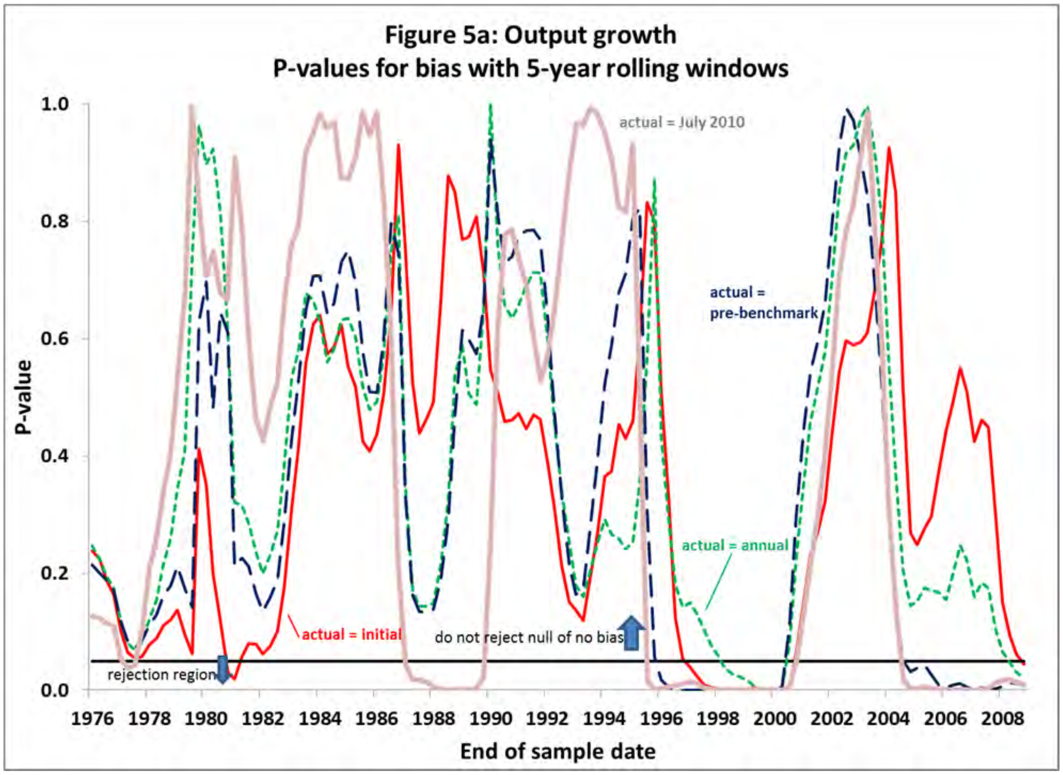


Figure 5a: Output growth
P-values for bias with 5-year rolling windows
1.0
0.8
0.6
0.4
0.2
0.0
P-value
actual = July 2010
actual =
pre-benchmark
actual = annual
actual = initial
do not reject null of no bias
rejection region
1976
1978
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
End of sample date

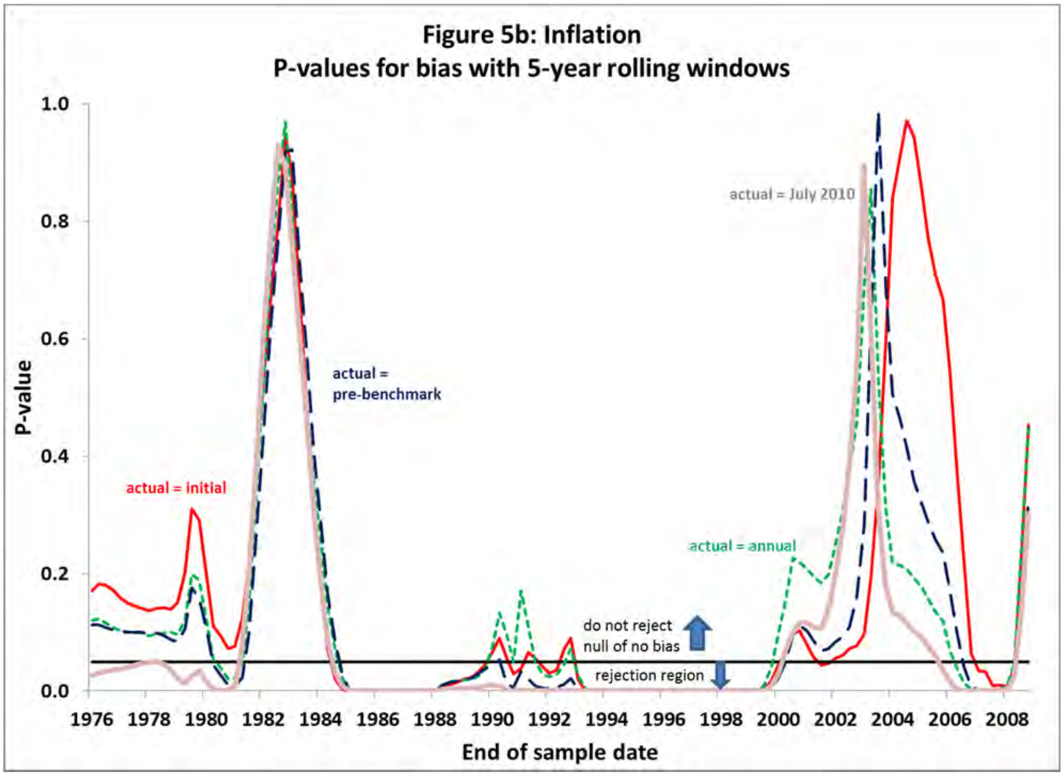


Figure 5b: Inflation
P-values for bias with 5-year rolling windows
1.0
0.8
0.6
0.4
0.2
0.0
P-value
actual = July 2010
actual =
pre-benchmark
actual = initial
actual = annual
do not reject
null of no bias
rejection region
1976
1978
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
End of sample date

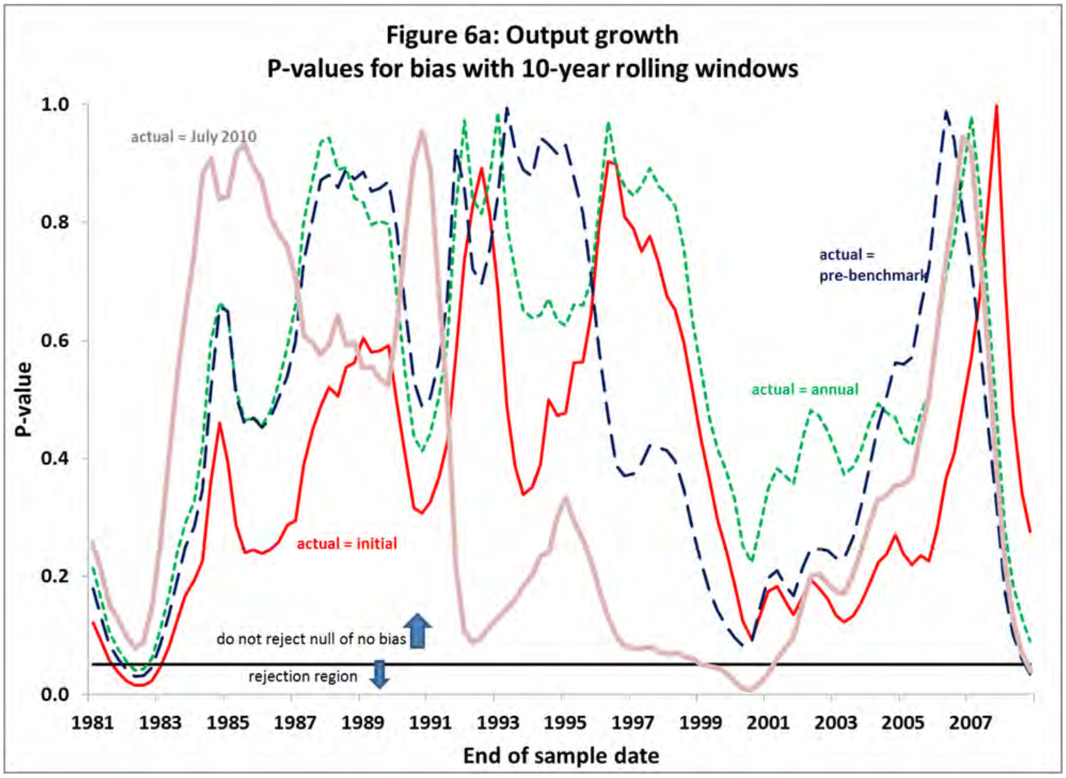


Figure 6a: Output growth
P-values for bias with 10-year rolling windows

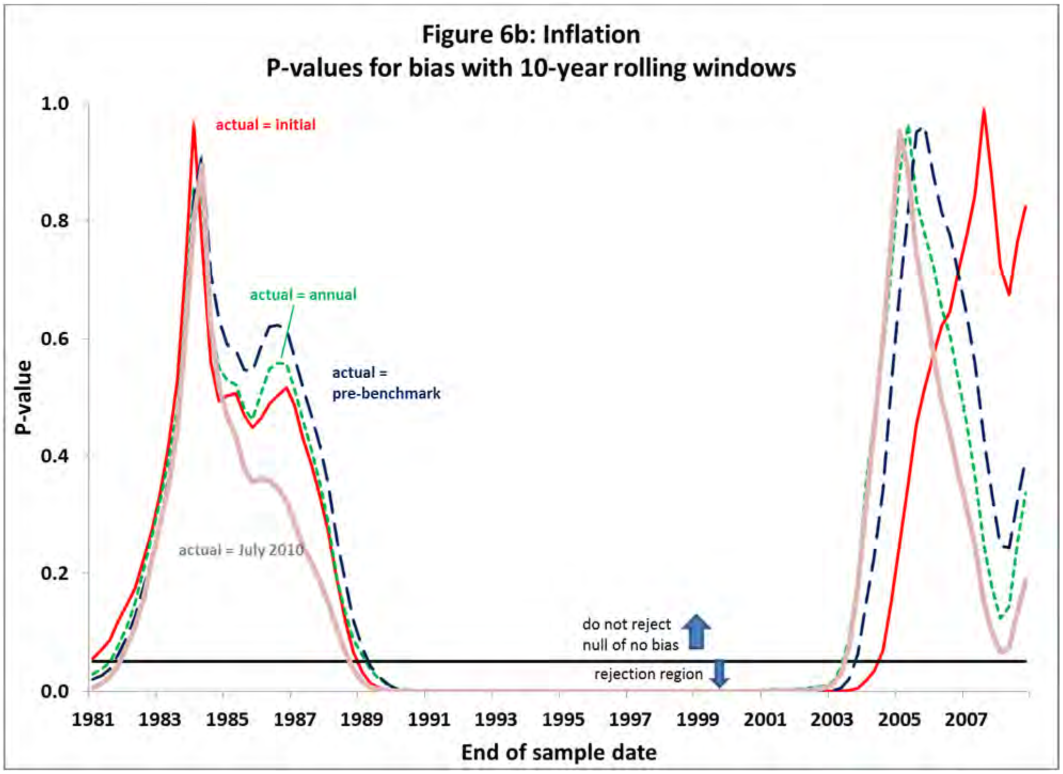


Figure 6b: Inflation
P-values for bias with 10-year rolling windows

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 choice of variable used as actual, especially for output growth.

An alternative method of testing for the optimality of forecasts in rolling samples was recently developed by Rossi and Sekhposyan (2011). They develop a general test for forecast optimality that is robust to the presence of instabilities, which are suggested by this paper’s results in Figures 3-6. 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.

Figures 7 and 8 show the results of the fluctuation optimality tests. In Figure 7, we examine forecasts of output growth, with part a showing 5-year rolling windows and part b showing 10-year rolling windows. Similarly, Figure 8 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-Sekhposyan (2011), Table 1b.

The results for output growth in Figures 7a and 7b are mixed. For the 5-year rolling forecasts shown in Figure 7a, there is evidence of nonoptimality in the SPF forecasts because samples ending in the second half of the 1990s and early 2000s 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 Figure 7b, however, the test value never exceeds the critical value, suggesting that the output growth forecasts from the SPF are not suboptimal.

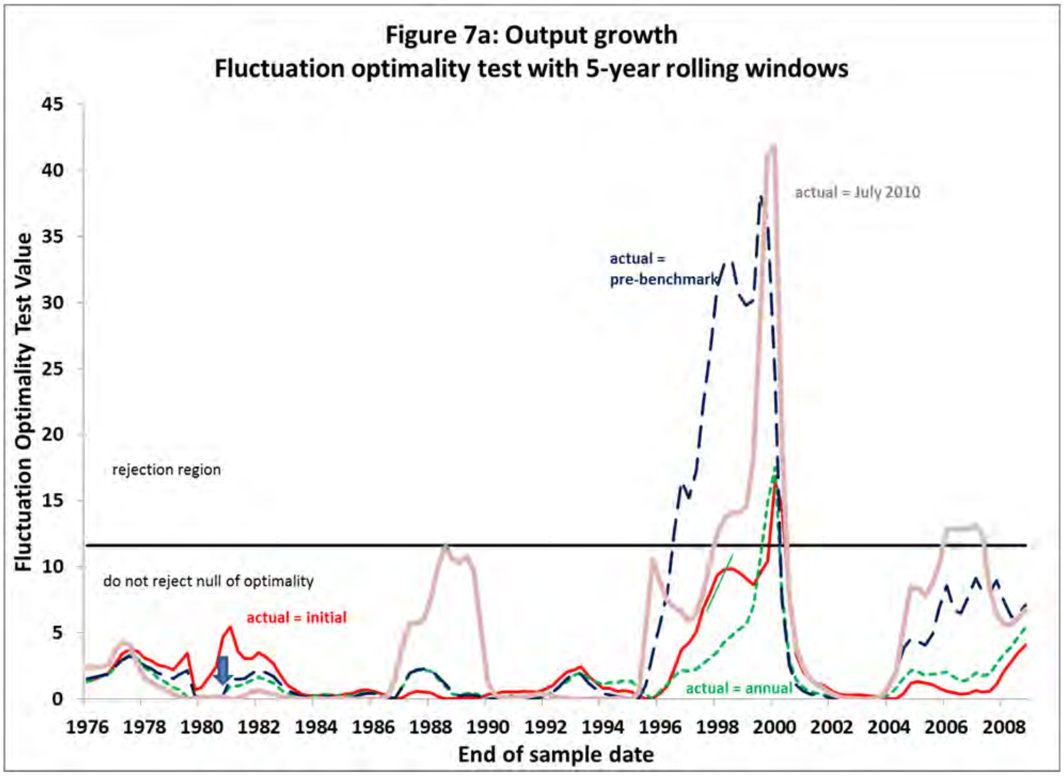


Figure 7a: Output growth
Fluctuation optimality test with 5-year rolling windows

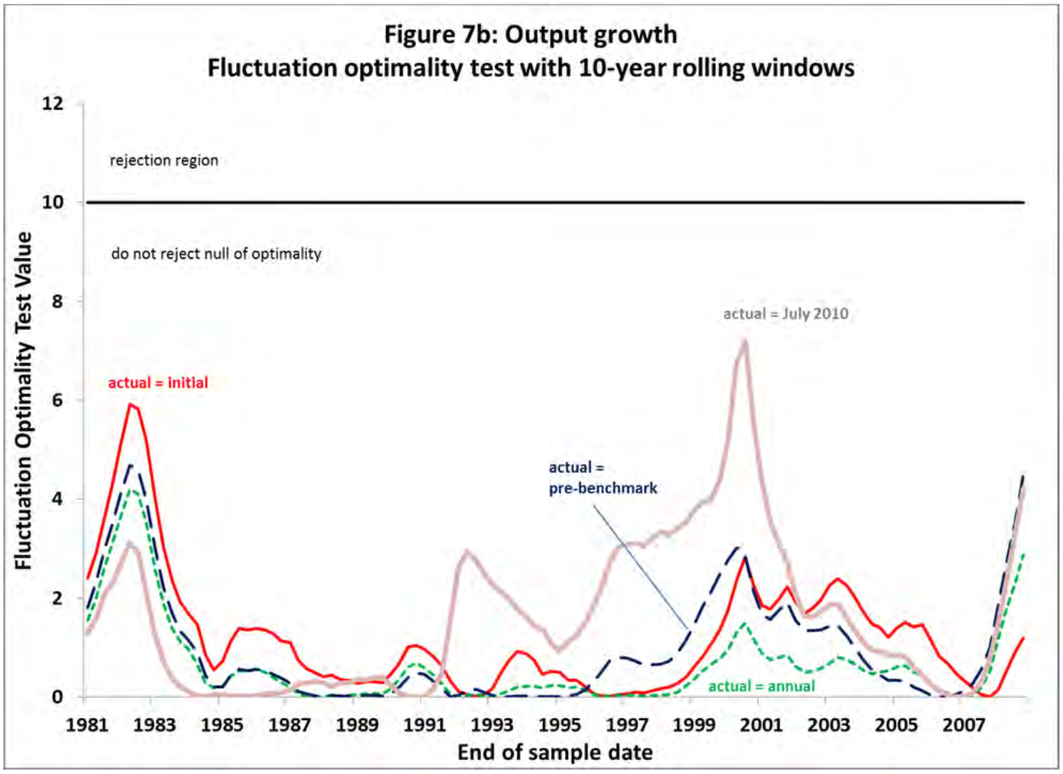
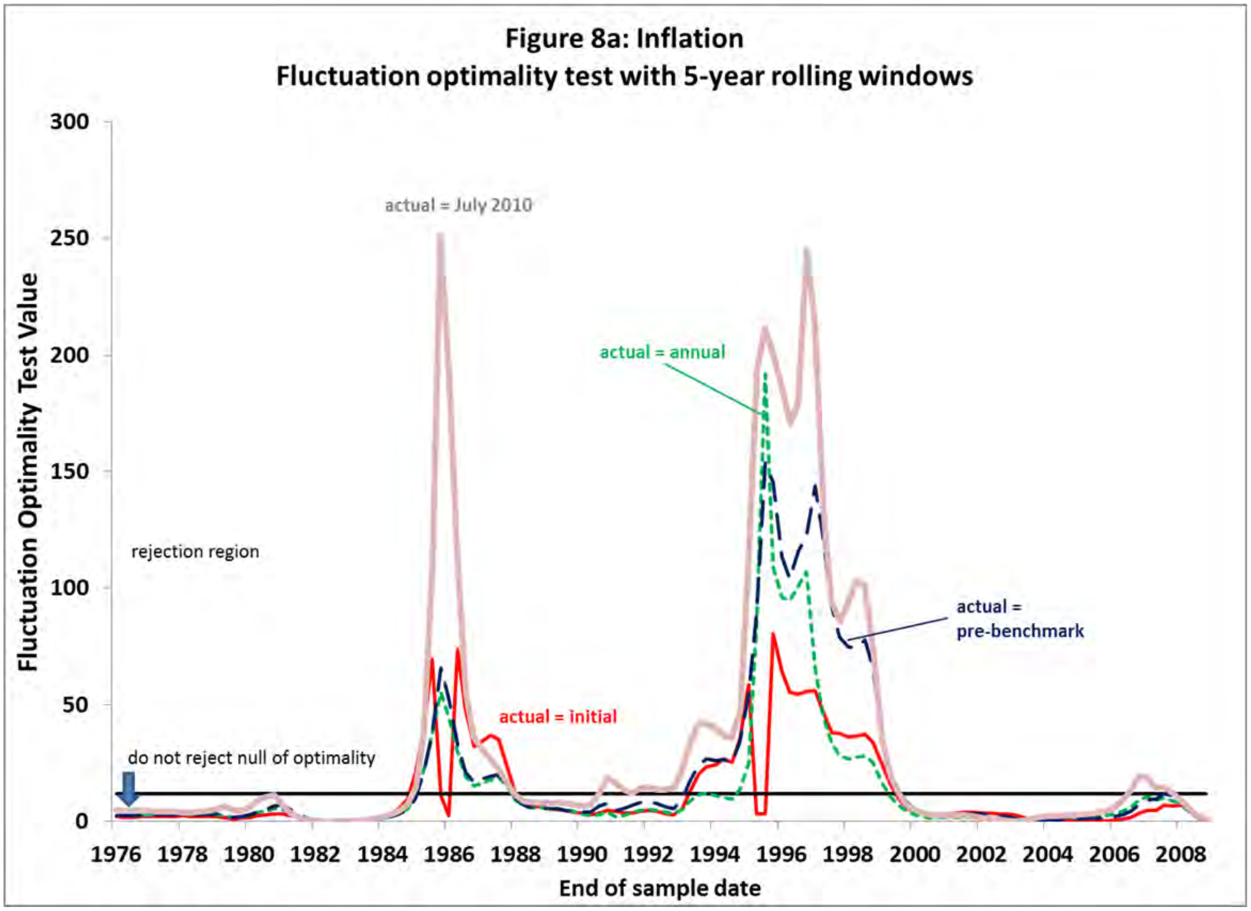
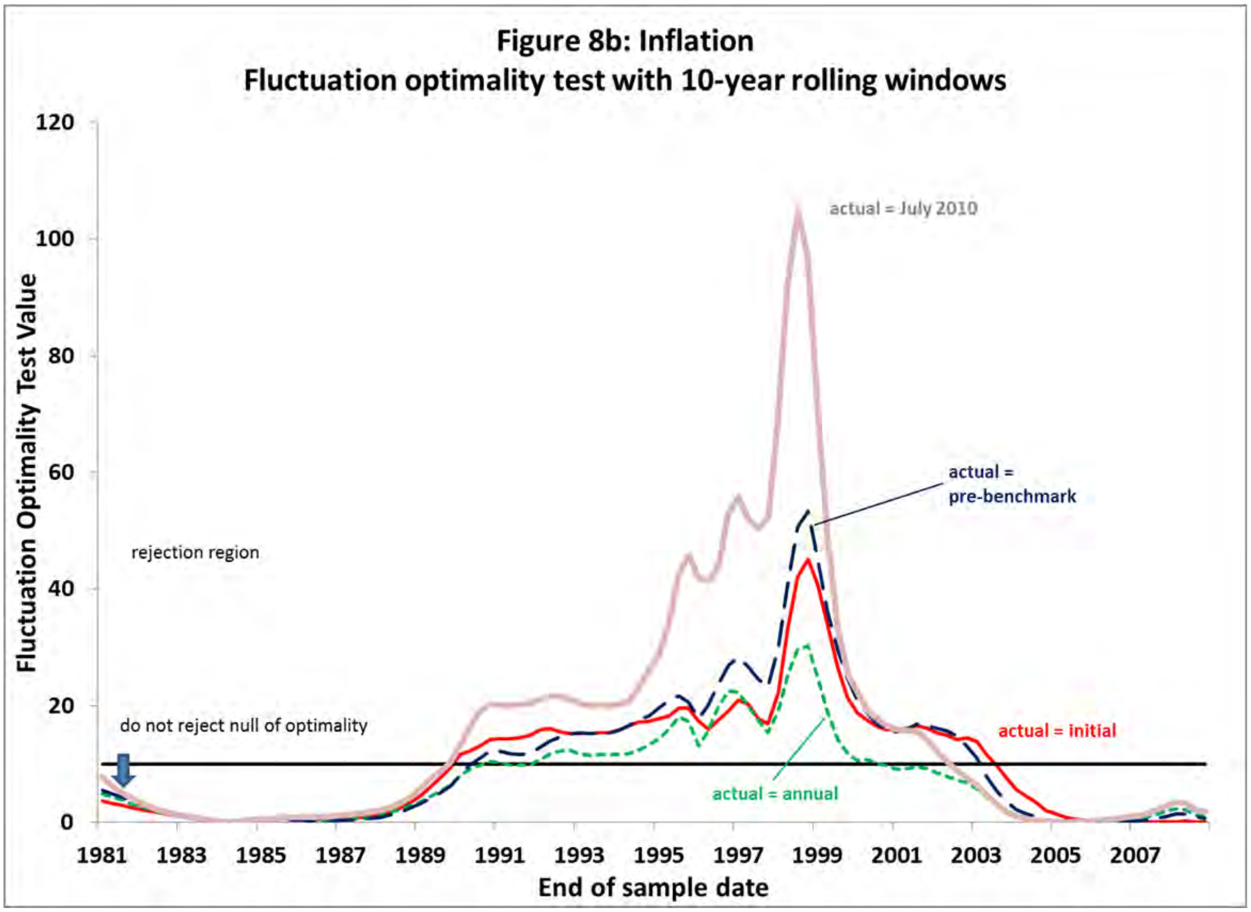


Figure 7b: Output growth
Fluctuation optimality test with 10-year rolling windows

For inflation forecasts, the fluctuation optimality tests are much worse, with test statistics that are very large compared with the critical value. In Figure 8a, illustrating inflation forecasts with 5-year rolling windows, subsamples that end in the mid-1980s or in the second half of the 1980s show very large test values far above the critical value. In Figure 8b, the inflation forecasts with 10-year rolling windows don't appear as bad, but samples ending in the 1990s and early 2000s show rejections. Thus, consistent with our other evidence, forecasts of inflation seem subject to instabilities that cause the forecasts to be suboptimal.





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.

To improve the forecasts, we estimate the bias using the initial release of the data to determine the forecast error on the left-hand side of equation (2), then create a new and improved forecast from the survey forecast:

$$\hat{\pi}_t = \hat{\alpha} + \pi_t^f, \qquad (3)$$

where $\hat{\pi}_t$ is the new and improved forecast and where 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 initial release of the data to determine the forecast error; we will modify that choice later. 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 2. The rows of the tables show alternative experiments, described below. The first column of numbers shows the root-mean-squared forecast error (RMSFE) for the original survey forecasts. The last three columns of the tables show alternative sample structures for estimating the bias: using the full sample, using 5-year rolling windows, and using 10-year rolling windows. The first number in each cell shows the RMSFE, and the second number shows the *p*-value for the test of a significantly different RMSFE, based on the Diebold-Mariano (1995) test.

The first row in Table 2 labeled “Adjust every period” for both output growth and inflation shows the results of the basic experiment in which we use equation (3) to attempt to improve on the survey forecasts based on the estimated bias each period. In every case, the forecasts are worse as 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 significant.

Table 2: RMSFEs and P-values for Forecast Improvement Exercises

Method	Original survey	Full sample	5-year window	10-year window
Output Growth				
Adjust every period	1.61	1.81 / 0.19	1.78 / 0.32	1.84 / 0.15
Adjust when $\rho < 0.05$	1.61	1.60 / 0.34	1.86 / 0.96	1.92 / 0.32
Adjust when $\rho < 0.05$ with shrinkage	1.61	1.61 / 0.34	1.59 / 0.50	1.64 / 0.32
Inflation				
Adjust every period	0.91	1.18 / 0.09	1.01 / 0.43	1.09 / 0.33
Adjust when $\rho < 0.05$	0.91	0.91 / 1.00	0.89 / 0.92	0.85 / 0.43
Adjust when $\rho < 0.05$ with shrinkage	0.91	0.91 / 1.00	0.84 / 0.14	0.84 / 0.10

Part of the reason for the poor performance of these attempts at forecast improvement is that we are trying to use the estimated bias from equation (2) even in periods when the bias is not statistically significant. However, more likely someone estimating equation (2) in real time would adjust the forecast using equation (3) only if the bias from estimating equation (2) was statistically significantly different from zero. So, suppose we follow this strategy. We will apply equation (3) only in periods when the p-value shown in Figures 3a and 3b for the initial release is below 0.05. We do the same for 5-year rolling windows based on the results shown in Figures 5a and 5b, and for 10-year rolling windows based on the results shown in Figures 6a and 6b.

The results of this exercise are shown in the second row for each variable in Table 2, labeled “Adjust when ρ < 0.05.” Compared with the results in the first row, the results here are much more promising. For output growth in the full sample and for inflation for the full sample and for 5-year rolling windows, the RMSFE is lower when using equation (3) to improve the forecast when the p-value is below 0.05. However, in no case is the difference statistically significant; in fact, the biggest difference is less than 7 percent of the original RMSFE. So, although it is possible to use equation (3) to improve the forecast in some cases, the forecast improvement is quite 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 apply equation (3), but only adjust for the bias by a factor of one-half:

$$\pi_t^i = (0.5 \times \hat{\alpha}) + \pi_t^f. \tag{4}$$

Using equation (4) instead of equation (3) and again only applying this adjustment when the p-value is less than 0.05, we get the results shown in the third row for each variable in Table 2, labeled “Adjust when ρ < 0.05 with shrinkage.”

The results show that shrinkage generally helps, especially with the 5-year and 10-year rolling windows. But still there is no statistically significant improvement from adjusting for the bias. 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.

The exercises reported in Table 2 use the initial release of the data at each date to estimate the bias in equation (2). However, a more common procedure in practice is for a researcher to use the real-time data available 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. That is, suppose that at each vintage date *v*, a research gathers the most recent data from a current database and uses those data to estimate equation (2), ignoring real-time data-revision issues altogether. Based on those estimates, suppose the researcher were to use equation (3) or (4) to improve upon the survey results, as before. The results of this exercise are shown in Table 3.

Table 3: RMSFEs and P-values for Forecast Improvement Exercises Using Most-Recent Vintage Data Each Period

Method	Original survey	Full sample	5-year window	10-year window
Output Growth				
Adjust every period	1.61	1.78 / 0.08	1.75 / 0.45	1.77 / 0.15
Adjust when $\rho < 0.05$	1.61	1.77 / 0.09	1.64 / 0.88	1.77 / 0.15
Adjust when $\rho < 0.05$ with shrinkage	1.61	1.68 / 0.13	1.60 / 0.82	1.68 / 0.22
Inflation				
Adjust every period	0.91	1.32 / 0.01	1.14 / 0.27	1.12 / 0.27
Adjust when $\rho < 0.05$	0.91	1.29 / 0.03	1.07 / 0.21	1.10 / 0.31
Adjust when $\rho < 0.05$ with shrinkage	0.91	1.09 / 0.03	0.93 / 0.74	0.97 / 0.50

The procedure that is the basis for Table 3 is based on the methods most commonly used by researchers. It assumes that the researcher uses the most recent data vintage available in real time and runs a real-time forecasting exercise at each period. In only one case (with rolling 5-year windows using shrinkage to forecast output growth) is there any improvement in RMSFEs, and that improvement is trivial and not statistically significant. In all other cases, the RMSFE increases, and the full-sample results for inflation show a statistically significant increase in the RMSFE.

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 GDP 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 3b, 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 3b, 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 1970-1974, which is consistent with the erratic nature of rejections of unbiasedness in the 5-year windows shown in Figure 5b. Bonham and Dacy (1991) found support for unbiasedness in SPF inflation forecasts from 1970-1984 using latest-available data, also consistent with Figure 3b. 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 3b. 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 3b, although the graph doesn’t quite extend to 2002. 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 does not appear to be possible to improve on the survey forecasts in real time; 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.

REFERENCES

Aiolfi, Marco, Carlos Capistran, and Allan Timmermann. "Forecast Combinations." Working paper, 2010, UC-San Diego.

Ang, Andrew, Geert Bekaert, and Min Wei. "Do Macro Variables, Asset Markets, or Surveys Forecast Inflation Better?" *Journal of Monetary Economics* 54 (May 2007), pp. 1163–1212.

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

Carroll, Christopher D. "Macroeconomic Expectations of Households and Professional Forecasters." *Quarterly Journal of Economics* 118 (February 2003), pp. 269–298.

Croushore, Dean. "Introducing: The Survey of Professional Forecasters." Federal Reserve Bank of Philadelphia *Business Review* (November/December 1993), pp. 3–13.

Croushore, Dean. "An Evaluation of Inflation Forecasts from Surveys Using Real-Time Data." *B.E. Journal of Macroeconomics: Contributions* (volume 10, issue 1, article 10, 2010).

Croushore, Dean. "Frontiers of Real-Time Data Analysis." *Journal of Economic Literature* 49 (March 2011), pp. 72-100.

Croushore, Dean, and Tom Stark. "A Real-Time Data Set for Macroeconomists," *Journal of Econometrics* 105 (November 2001), pp. 111–130.

Diebold, Francis X., and Roberto S. Mariano. "Comparing Predictive Accuracy." *Journal of Business and Economic Statistics* 13 (July 1995), 253-263.

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

Giacomini, Raffaella, and Barbara Rossi. “Forecast Comparisons in Unstable Environments,” *Journal of Applied Econometrics* 25 (2010), pp. 595-620.

Hafer, R. W., and Scott E. Hein. “On the Accuracy of Time-Series, Interest Rate, and Survey Forecasts of Inflation.” *Journal of Business* 58 (October 1985), pp. 377–398.

Hansen, Lars-Peter, and Robert J. Hodrick. “Foreign Exchange Rates as Optimal Predictors of Future Spot Rates: An Econometric Analysis,” *Journal of Political Economy* 88 (October 1980), pp. 829-53.

Keane, Michael P., and David E. Runkle. “Testing the Rationality of Price Forecasts: New Evidence from Panel Data.” *American Economic Review* 80 (September 1990), pp. 714–735.

Mankiw, N. Gregory, Ricardo Reis, and Justin Wolfers. “Disagreement About Inflation Expectations.” NBER *Macroeconomics Annual 2003*, pp. 209–248.

Mankiw, N. Gregory, and Matthew D. Shapiro. “Do We Reject Too Often? Small Sample Bias in Tests of Rational Expectations Models,” *Economics Letters* 20 (1986) 139-145.

Mincer, Jacob A., and Victor Zarnowitz. “The Evaluation of Economic Forecasts.” In: Jacob Mincer, ed., *Economic Forecasts and Expectations* (New York: National Bureau of Economic Research, 1969.)

Newey, Whitney K., and Kenneth D. West. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica* 55 (May 1987), pp. 703-8.

Patton, Andrew J., and Allan Timmermann. "Forecast Rationality Tests Based on Multi-Horizon Bounds." Working paper, UC-San Diego, 2011.

Romer, Christina D. and David H. Romer. "Federal Reserve Information and the Behavior of Interest Rates." *American Economic Review* 90:3 (June 2000), pp. 429-457.

Rossi, Barbara. "Are Exchange Rates Really Random Walks? Some Evidence Robust to Parameter Instability." *Macroeconomic Dynamics* 10 (2006), pp. 20-38.

Rossi, Barbara, and Tatevik Sekhposyan. "Forecast Optimality Tests in the Presence of Instabilities." Working Paper, Duke University, August 2011.

Stock, James H., and Mark W. Watson. "Forecasting Output and Inflation: The Role of Asset Prices." *Journal of Economic Literature* 41 (Sept. 2003), pp. 788-829.

Su, Vincent, and Josephine Su. "An Evaluation of ASA/NBER Business Outlook Survey Forecasts." *Explorations in Economic Research* 2 (Fall 1975), pp. 588–618.

Zarnowitz, Victor. "Rational Expectations and Macroeconomic Forecasts." *Journal of Business and Economic Statistics* 3 (October 1985), pp. 293–311.