AN EVALUATION OF INFLATION FORECASTS FROM SURVEYS USING REAL-TIME DATA

Dean Croushore

Associate Professor of Economics and Rigsby Fellow
University of Richmond

Visiting Scholar
Federal Reserve Bank of Philadelphia

December 2008

I thank two referees and the editor for suggestions, as well as Kundan Kishor, Bryan Campbell, Sean Campbell, and discussants and participants at the 2005 Workshop on Macroeconomic Forecasting, Analysis, and Policy Design with Data Revisions in Montreal, the 2005 Southern Economic Association meetings, the 2006 American Economic Association meetings, the 2006 Econometric Society Summer meetings, the University of Richmond economics seminar, and the 2007 International Symposium on Forecasting for useful comments on this paper. Amanda Smith provided outstanding research assistance on this project. 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.

AN EVALUATION OF INFLATION FORECASTS FROM SURVEYS USING REAL-TIME DATA

ABSTRACT

This paper carries out the task of evaluating inflation forecasts from the Livingston Survey and the Survey of Professional Forecasters, using the Real-Time Data Set for Macroeconomists as a source of real-time data. We examine the magnitude and patterns of revisions to the inflation rate based on the output price index. We then run tests on the forecasts from the surveys to see how good they are. We find that there are several episodes in which forecasters made persistent forecast errors, but the episodes are so short that by the time they can be identified, they have nearly disappeared. Thus, improving on the survey forecasts seems to be very difficult in real time, and the attempt to do so leads to increased forecast errors.

AN EVALUATION OF INFLATION FORECASTS FROM SURVEYS USING REAL-TIME DATA

As part of the research program into rational expectations in the early 1980s, economists tested the forecasts of consumer price inflation from surveys (the Livingston Survey and the Survey of Professional Forecasters) and found a disturbing result: the forecasts exhibited bias and inefficiency. If macroeconomic forecasters had rational expectations, the forecast errors should have had much better properties; instead, the forecasters appeared to make systematic errors. Researchers concluded that perhaps macroeconomic forecasters were irrational or perhaps the surveys were poor measures of inflation expectations. Simple univariate time-series models had lower forecast errors than the surveys of professional forecasters. The major consequence was that forecast surveys developed a poor reputation that has continued to this day. Many researchers ignore forecast surveys as a source of data on people's expectations.[1]

But perhaps the researchers in the early 1980s were hasty in their condemnation of the surveys. If the researchers were correct, then it should have been a simple task to use their empirical results and provide new and improved forecasts. The question is, were their results special to the data sample of the time? Also, were their results possibly an artifact of the data they were using?

[1]See Maddala (1991) and Thomas (1999) for literature reviews and Carroll (2003) for a comparison of expectations formation between households and professional forecasters.

Most studies have focused on bias tests for inflation forecasts by looking at the consumer price index.[2] But the CPI is not the best measure of inflation because of index construction problems, as described in the Boskin (1996) commission report. Better measures of trend inflation come from other variables, such as the GDP deflator. But forecasts of the inflation rate using the GDP deflator are more difficult to evaluate because the past data are revised. The sample period in which most of the earlier tests were performed was a time with numerous shocks to relative prices, which were only slowly reflected in the GDP weights, thus leading to a delayed pattern of revisions. As a result, a real-time analysis of the data is paramount.

Only a few studies have examined the forecasts of the GDP inflation rate as we do here. A number of papers investigate bias in the inflation forecasts but suffer from a failure to account for the nature of real-time data, including Dua and Ray (1992), Hafer and Hein (1985), Rhim et al. (1994), and Vanderhoff (1984). The most notable study that accounts for data revisions and finds bias in the forecasts is Zarnowitz (1985), who finds bias in the SPF forecasts from 1968 to 1979, using pre-benchmark vintages as actuals. By contrast, Keane and Runkle (1990) find evidence against bias using individual data, rather than the survey average data that Zarnowitz used. Following up on Zarnowitz, Baghestani and Kianian (1993) found the SPF GDP inflation forecasts biased from 1981 to 1991, though they used the initial release as actuals. Davies (2006) uses a sophisticated

[2] For example, the recent paper by Ang et al. (2007) provides important evidence that survey forecasts are superior to many other forecasting methods, but is based mostly on the consumer price index.

framework accounting for differences across forecasters, horizons, and variables, and finds that about one-fourth of the forecasters exhibit bias in their forecasts.

The only other papers that use real-time data to test for bias and inefficiency in the inflation forecast surveys are Capistrán and Timmermann (2006) and Davies (2006). Both find bias and inefficiency by individual forecasters, rather than the average across forecasters. Capistrán and Timmermann find that although individuals' forecasts are biased, the biases offset so that the mean forecast is not biased. The results in this paper are similar to the results for the average inflation forecast across forecasters found by Capistrán and Timmermann, but this paper uses a different set of tests, shows how real-time data may affect forecast evaluation, and examines alternative sample periods.

This paper carries out the task of evaluating inflation forecasts from the Livingston Survey and the Survey of Professional Forecasters, using real-time data. We begin by examining the magnitude and patterns of revisions to the GDP inflation rate. We then run tests for bias on the forecasts from the surveys to see how good they are, and attempt to use real-time information to improve on the survey forecasts. We find that there are several episodes in which forecasters made persistent forecast errors, but the episodes are so short that by the time they can be identified, they have nearly disappeared. Thus, improving on the survey forecasts seems to be very difficult in real time, and the attempt to do so leads to increased forecast errors. We also investigate whether simple time-series models can do better than the survey forecasts and find that they do not, except in certain sub-sample periods. As Ang et al. (2007) found, the forecast surveys are difficult to beat in real time.

THE DATA

In examining data on inflation and forecasts of inflation, we must account for the noise in high frequency measures of the data. Analysts and policymakers typically do not care about monthly or quarterly movements of inflation, which depend on short-term shocks, but usually analyze it over longer periods, such as one year.[3] Because forecasts are often taken at such a frequency, the focus of this paper is on inflation and inflation forecasts measured over (roughly) a one-year time span.

The Federal Reserve Bank of Philadelphia's Real-Time Data Set for Macroeconomists collects snapshots of numerous macroeconomic time series data once each quarter since November 1965.[4] Data within any vintage of the data set can be used to show precisely what official data were available to a forecaster at any given date. The GDP deflator is one of the variables included within the data set. The timing of the vintages is as of the middle day of the middle month of each quarter.

The only two surveys that span the period from the 1970s to today with forecasts for the GDP deflator are the Livingston Survey and the Survey of Professional Forecasters. The Livingston Survey, which began in the 1940s, collects its forecasts from a wide array of economists, not all of whom have forecasting as their main job. The Survey of Professional Forecasters (SPF), which was known as the ASA-NBER Survey

---

[3] Bryan and Cecchetti (1994) provide a cogent description of the noise in inflation data.

[4] Croushore and Stark (2001) describe the structure of the Real-Time Data Set for Macroeconomists and evaluate data revisions to some variables. Croushore (2006) shows how data revisions affect forecasts, while Croushore and Stark (2003) illustrate how data revisions have influenced major macroeconomic research studies. See the Philadelphia Fed's website for the data set at: http://www.philadelphiafed.org/research-and-data/real-time-center/real-time-data.

from 1968 to 1990 before it was taken over by the Federal Reserve Bank of Philadelphia, collects its forecasts from economists for whom forecasting is a major part of their jobs.

The Livingston Survey collects economists' forecasts of levels of real output and nominal output (GNP until 1991, GDP since 1992).[5] From these forecasts, we can calculate the implicit forecasts of inflation in the GNP or GDP deflator. Survey forms are sent out about mid-May and mid-November each year and are due back in early June and December. Because the first-quarter values of real output and nominal output are revised in late May each year, we assume that the forecasters knew the revised numbers before making their forecast, so we include those data in our real-time data set. Similarly, we assume the forecasters know the value for the third quarter that is released in late November before making their forecasts. Because the survey calls for forecasts through the second quarter of the following year (for surveys due in June) and the fourth quarter of the following year (for surveys due in December), the forecasters are really making five-quarter-ahead forecasts. Although the survey itself began in 1946 and forecasts for nominal output have been part of the survey since it began, forecasts for the level of real output did not begin until June 1971. So we begin our sample with that survey. Our sample ends with the survey made in December 2006 because that is the last survey whose one-year-ahead forecasts we can evaluate (as of February 2008 when the data for

[5] See Croushore (1997) for details on the Livingston Survey. The survey's data are all available online at: http://www.philadelphiafed.org/research-and-data/real-time-center/livingston-survey.

this draft of the paper were collected). To avoid idiosyncratic movements in the forecasts, we examine the median forecast across the forecasters.[6]

The Survey of Professional Forecasters collected forecasts of the GNP deflator from 1968 to 1991, the GDP deflator from 1992 to 1995, and the GDP price index since 1996.[7] The GNP deflator, GDP deflator, and GDP price index behave quite similarly, and there is no apparent break in the inflation series generated from the three different price indexes in either 1992 or 1996. From these forecasts, we can calculate the implicit forecasts of inflation. Survey forms are sent out four times a year after the advance release of the national income and product accounts in late January, April, July, and October and are due back before the data are revised in February, May, August, and November. The survey calls for forecasts for each quarter for the current and following four quarters, so we can construct an exact four-quarter-ahead forecast. The timing can be seen in the following example: in late January 2004, the national income account data are released and the forecasters know the values of the GDP price index from 1947:Q1 to 2003:Q4. They forecast levels of the GDP price index for 2004:Q1, 2004:Q2, 2004:Q3, 2004:Q4, and 2005:Q1. We examine their one-year-ahead forecasts based on their forecast for 2005:Q1 relative to their forecast for 2004:Q1. Thus, the forecasts span a one-year (four-quarter) period, though it may be relevant to note that the end of their

[6] A recent paper by Capistrán-Timmermann (2007) supports the use of the median forecast. They find that equal-weighting of survey forecasts leads to lower root-mean-squared forecast errors than other methods of forecast combination. For the Survey of Professional Forecasters and Livingston Survey, the mean forecast is very close to the median forecast.

[7] See Croushore (1993) for more on the SPF. The survey forecasts can be found online at www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters. Forecasters in the SPF are given the opportunity to forecast data revisions, but do not do so very often.

forecast horizon (2005:Q1) is five quarters after the last date for which they observe a realization (2003:Q4). Although the survey itself began in 1968, the early forecasts for the GNP deflator were rounded to the nearest whole number, which causes the forecasts to be quite erratic in the early years of the survey. Because of this, and to analyze the Livingston Survey and SPF forecasts on the same sample period, we look at the SPF forecasts made between 1971:Q1 and 2006:Q4. Our sample ends with the surveys made in 2006:Q4, because that is the last survey whose one-year-ahead forecasts we can evaluate, given realized data through 2007:Q4 (the latest available data when this draft was written). As with the Livingston Survey, to avoid idiosyncratic movements in the forecasts, we examine the median forecast across the forecasters.

Let us begin our data analysis by looking at plots of the forecasts over time and some measures of the actual values of the GDP inflation rate. For the Livingston Survey, we plot forecasts and actuals based on latest available data (as of February 2008) in Figure 1 from 1971:H1 to 2006:H2. In our date notation, "H" means "half year"; so, for example, the survey from 2006:H2 means the survey made in the second half of 2006, which was released in December 2006. The dates on the horizontal axis represent the dates at which a forecast was made. The corresponding "forecast" point is the forecast for the five-quarter period from the first quarter of the current year to the second quarter of the following year for June surveys, and from the third quarter of the current year to the fourth quarter of the following year for December surveys. For example, the data points shown in the upper panel for 2006:H2 (the data points furthest to the right on each line) are: (1) the forecast from the December 2006 Livingston Survey for the GDP inflation rate from 2006:Q3 to 2007:Q4; and (2) the actual GDP inflation rate based on latest

available data (dated February 2008) from 2006:Q3 to 2007:Q4. In the lower panel of Figure 1, the forecast error is shown (defined as the actual value minus the forecast).

Figure 1
Median Inflation Forecast from Livingston Survey

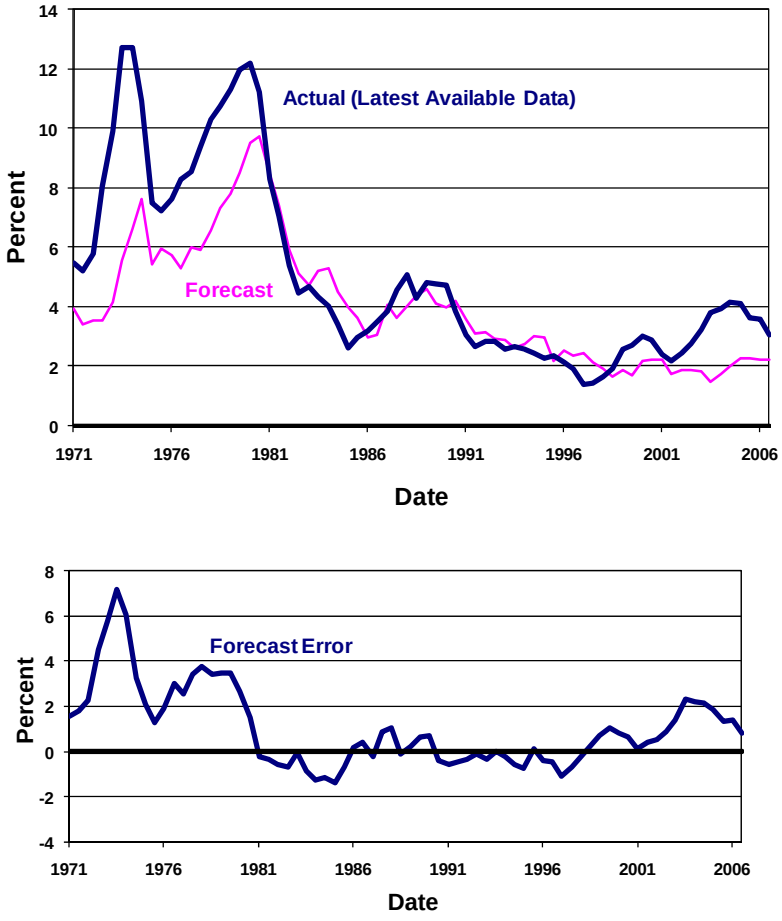
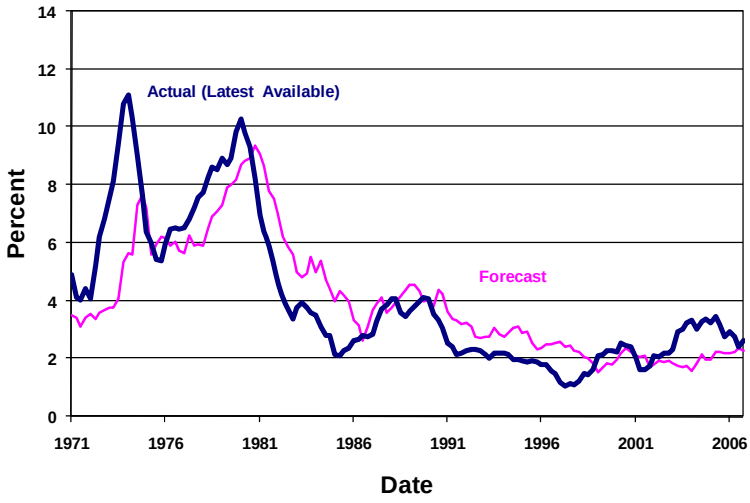
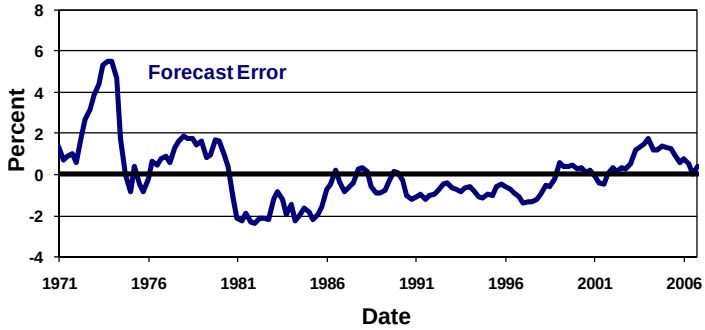


Figure 2 shows a similar plot for the Survey of Professional Forecasters. For the SPF, we plot forecasts and actuals based on latest available data (as of February 2008) in Figure 2 from 1971:Q1 to 2006:Q4. As in Figure 1, the dates on the horizontal axis represent the dates at which a forecast was made. The corresponding "forecast" point is the forecast for the four-quarter period from the date shown on the horizontal axis; for

example, the data points shown in the upper panel for 2006:Q4 (the data points furthest to the right on each line) are: (1) the forecast from the November 2006 SPF for the GDP inflation rate from 2006:Q4 to 2007:Q4; and (2) the actual GDP inflation rate based on latest available data (dated February 2008) from 2006:Q4 to 2007:Q4. In our date notation, "Q" means "quarter"; so, for example, the survey from 2006:Q4 means the survey made in the fourth quarter of 2006, which was released in November 2006. In the lower panel of Figure 2, the forecast error is shown (defined as the actual value minus the forecast).

Figure 2
Median Inflation Forecast from SPF





The figures for both the Livingston Survey and the SPF have three features in common: (1) inflation rose much higher than the forecasters thought it would in the 1970s (more so in the Livingston Survey than in the SPF); (2) the forecasters were slow to reduce expected inflation in the early 1980s and their forecast errors were negative for a time (more so in the SPF than in the Livingston Survey); and (3) SPF forecast errors were persistently negative in the 1990s (meaning that the forecasters thought inflation would be higher than it turned out to be), averaging -0.7 percentage points (with inflation averaging 2.2%).

REAL-TIME DATA ISSUES

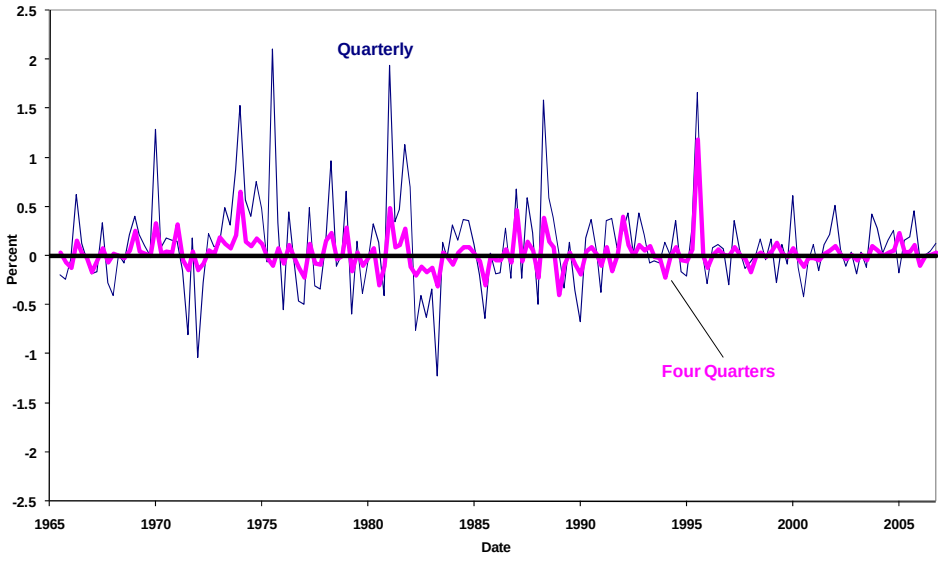
Before we examine the quality of the forecasts, we must tackle the difficult issue of what to use as actuals for calculating forecast errors. In the discussion above, we based our analysis solely on the latest available data (as of February 2008), which is what is typically done in the forecasting literature. But forecasters are quite unlikely to have anticipated the extent of data revisions to the price index that would not occur for many years in the future. More likely, they made their forecasts anticipating the same methods of data construction being used contemporaneously by the government statistical agencies.

How big are the revisions to the data on the price index for output? In the real-time data set we can consider a variety of actuals, including the value recorded one quarter after the initial release, the value recorded one year after the initial release, the last value recorded before a new benchmark revision occurs (a concept that maintains a consistent method by which the government calculates growth rates, including the same

base year) about every 5 years, and the value recorded in the latest available data (as of February 2008). How different are these alternative concepts of actuals? And how large are the consequent data revisions? In this paper, we choose to use the value recorded one quarter after the initial release. The working paper version of this paper considered other possibilities; see Croushore (2006) for those variations.

Revisions from initial release to one quarter later can be large, as Figure 3 shows.[8] The figure shows the relative size of the revisions for both quarterly data (so you can observe when revisions to the one-quarter GDP inflation rate occur) and four-quarter data (our central object of interest). A number of revisions in the quarterly inflation rate

Figure 3
Revisions to Inflation Rate
Initial to One Quarter Later

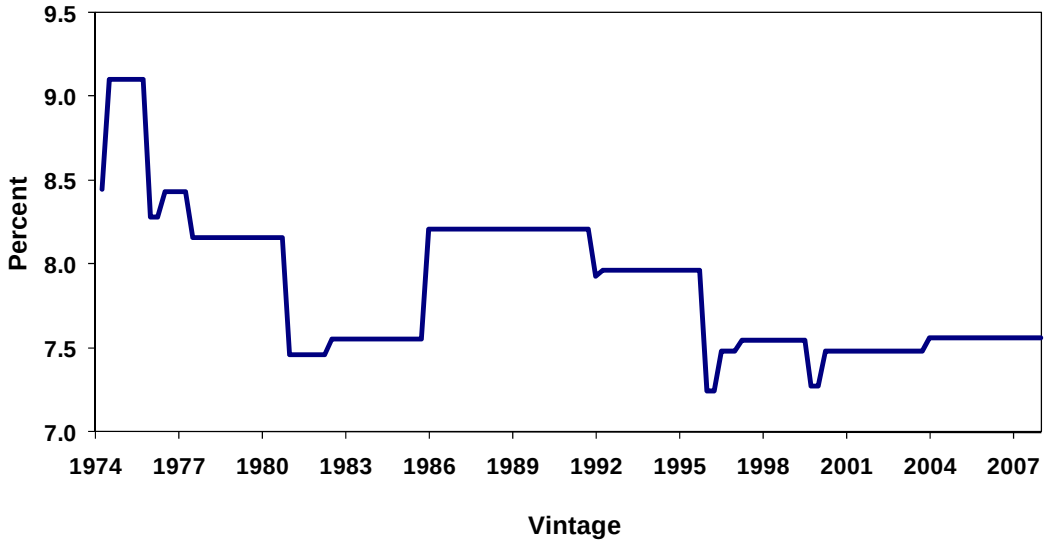


[8] Aruoba (2008) finds that the annual inflation rate in the output deflator has a nonzero mean revision, though his concept of revision is slightly different from the one we use here, since he does not look at revisions caused by benchmark revisions.

exceed 1 percent, and revisions to the four-quarter GDP inflation rate sometimes exceed 0.5 percent.

For any observation date, it is possible to track the GDP inflation rate as it is successively revised across all the vintages since its initial release. This is done in Figure 4 for the GDP inflation rate between 1973:Q1 and 1974:Q1. The GDP inflation rate for this period was initially released as 8.4 percent in the vintage of May 1974 and revised substantially over the next 30 years of data vintages. With so many fluctuations in the measure of the GDP inflation rate, it is clear that the result of any statistical method that evaluates a forecast made in 1973:Q1 for the coming year is going to depend significantly on what is chosen to serve as the actual value of the variable.

Figure 4
Inflation Rate from 1973Q1 to 1974Q1
(as viewed from the perspective of 136 different vintages)



Even over longer forecast horizons, revisions to the GDP inflation rate are significant. To see this, consider the average GDP inflation rate over five-year intervals, measured at different vintage dates. In Table 1, we consider vintage dates for every pre-benchmark vintage and show the five-year average GDP inflation rate. Across vintages, the five-year average GDP inflation rate changes by as much as 0.7 percentage point. Thus, the value of the GDP inflation rate for substantial periods of time is not precisely measured.

ARE THE FORECASTS BIASED?

In the literature on testing forecasts for accuracy, a key test is one to see if the forecasts are biased. A forecast is biased if forecasts differ systematically from their realized values. A glance at the early years shown in Figures 1 and 2 suggests that the forecasts might be biased. Because many of the studies of bias in the survey data were undertaken in the early 1980s, during the period in which the theory of rational expectations was being tested empirically, it is clear why the tests suggested that the survey forecasts were not rational. Scatter plots of the data from both surveys allow us to eyeball the bias in the surveys, as shown in Figure 5. In the period from 1971 to 1981, there is a clear tendency in both surveys for the forecasts to be too low (points above and to the left of the 45-degree line) relative to actuals. After that, however, the forecasts are much better in both surveys, with a slight tendency in the SPF for the forecasts of inflation to be too high.

Table 1
Average Inflation Rate Over Five Years
For Pre-Benchmark Vintages
Annualized percentage points

Vintage Year:	1975	1980	1985	1991	1995	1999	2003	2008
Period								
49Q4 to 54Q4	2.6	2.7	2.7	2.5	2.4	2.6	2.5	2.6
54Q4 to 59Q4	2.6	2.6	2.6	2.9	2.9	2.4	2.5	2.5
59Q4 to 64Q4	1.4	1.5	1.5	1.6	1.6	1.3	1.3	1.3
64Q4 to 69Q4	3.6	3.9	3.9	4.1	4.1	3.7	3.7	3.7
69Q4 to 74Q4	6.3	6.5	6.2	6.8	6.5	6.3	6.3	6.3
74Q4 to 79Q4	NA	7.1	7.0	7.5	7.7	7.2	7.1	7.1
79Q4 to 84Q4	NA	NA	6.1	6.1	6.4	6.2	6.0	6.0
84Q4 to 89Q4	NA	NA	NA	3.3	3.6	3.4	3.1	3.0
89Q4 to 94Q4	NA	NA	NA	NA	2.9	3.1	2.8	2.7
94Q4 to 99Q4	NA	NA	NA	NA	NA	NA	1.7	1.6
99Q4 to 04Q4	NA	NA	NA	NA	NA	NA	NA	2.4

This table shows the inflation rates over the five year periods shown in the first column for each pre-benchmark vintage shown in the column header.

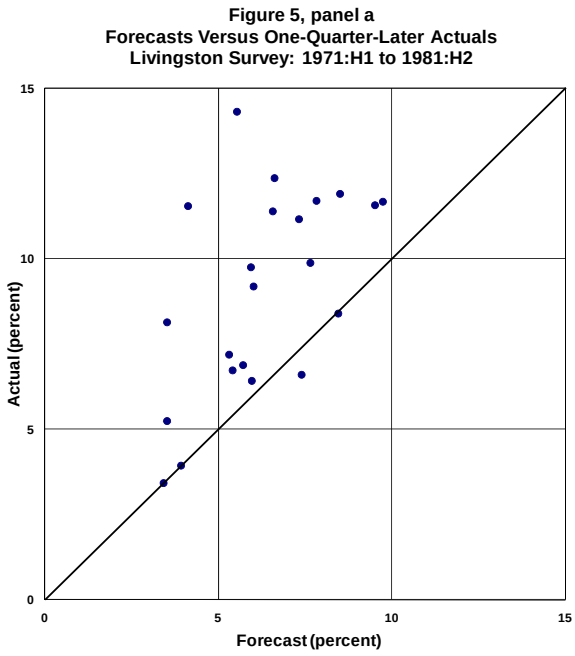
To examine the bias more formally, researchers often run a Mincer-Zarnowitz (1969) test, based on the regression:

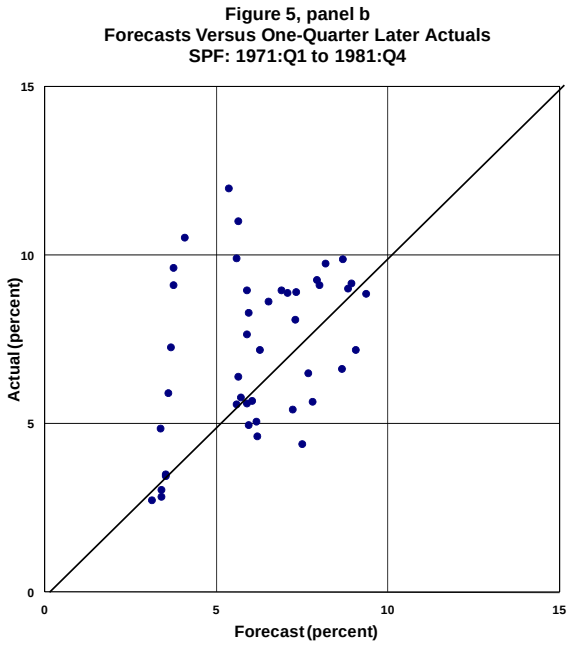
$$\pi_t = \alpha + \beta \pi_t^f + \varepsilon_t , \tag{1}$$

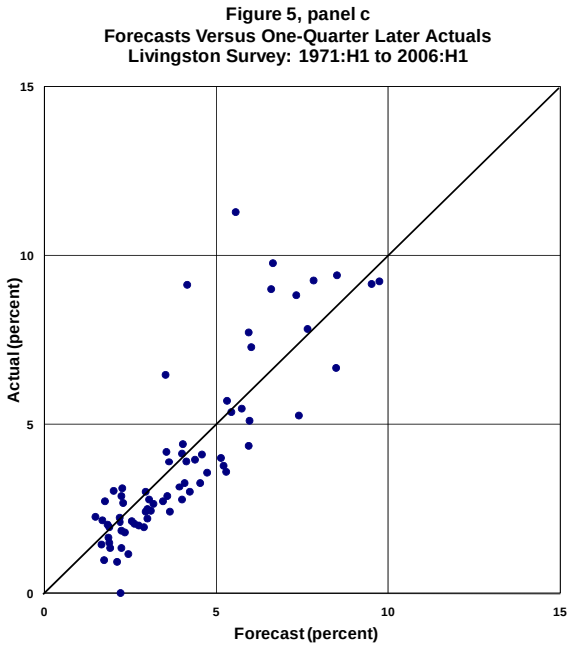
where π_t is the actual inflation rate and π_t^f is the forecast at each date *t*. If the forecasts are not biased, we should estimate $\hat{\alpha} = 0$ and $\hat{\beta} = 1$, as first suggested by Theil (1966). Webb (1987), however, has challenged this view of bias, arguing that even if we reject this joint hypothesis, data revisions, coefficients that change over time, and peso-type problems may prevent someone from using the results of Equation (1) to make better forecasts. More importantly, Mankiw-Shapiro (1986) argue that because of the autoregressive nature of the variable on the right-hand side of the equation, there is a small-sample bias that tends to reject the null of rationality too often. To remedy this, we modify the regression to:

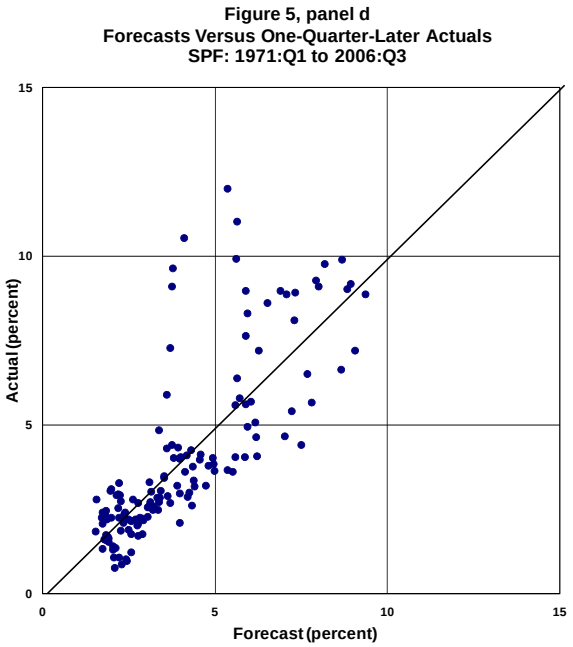
$$e_t = \pi_t - \pi_t^f = \alpha + \varepsilon_t , \tag{2}$$

so that this is effectively a test for a zero mean of the forecast error.









In testing forecasts over a four-quarter (SPF) or five-quarter (Livingston) horizon, we face the issue of overlapping observations. Because the forecasts span a longer period

than the sampling frequency, any shock affects the actuals for several consecutive periods. For example, an oil-price shock in 1973:Q2 affects any measurement of actuals that includes that quarter and therefore the forecast errors for forecasts made over the period, including 1973:Q2. For the SPF, this means that the forecast errors from surveys taken in 1972:Q2, 1972:Q3, 1972:Q4, 1973:Q1, and 1973:Q2 are all correlated; for the Livingston Survey, correlation occurs among forecast errors from surveys taken in 1972:H1, 1972:H2, and 1973:H1. To allow for these overlapping observations, we must either cut the SPF sample into five pieces (taking every fifth observation) and the Livingston Survey into three pieces, or adjust the covariance matrix using methods suggested by Brown and Maital (1981), using the method of Hansen and Hodrick (1980), perhaps as modified by Newey and West (1987) to guarantee a positive definite covariance matrix. I will report only the latter results, but they are largely consistent with results from cutting the survey into non-overlapping observations.[9]

When we run the regression in equation (2) over the entire sample period, using one-quarter later data as actuals, we find no evidence of bias, as Table 2 shows. The mean values of the forecast errors are close to zero and the p-values are far from 0.05, so we do not reject the null hypothesis of mean-zero forecast errors. However, if we split the sample in pieces, we can see that in some sub-samples, there is evidence of bias. To investigate this issue, we cut the sample at every date between the first quarter of 1978 and the last quarter of 1998.

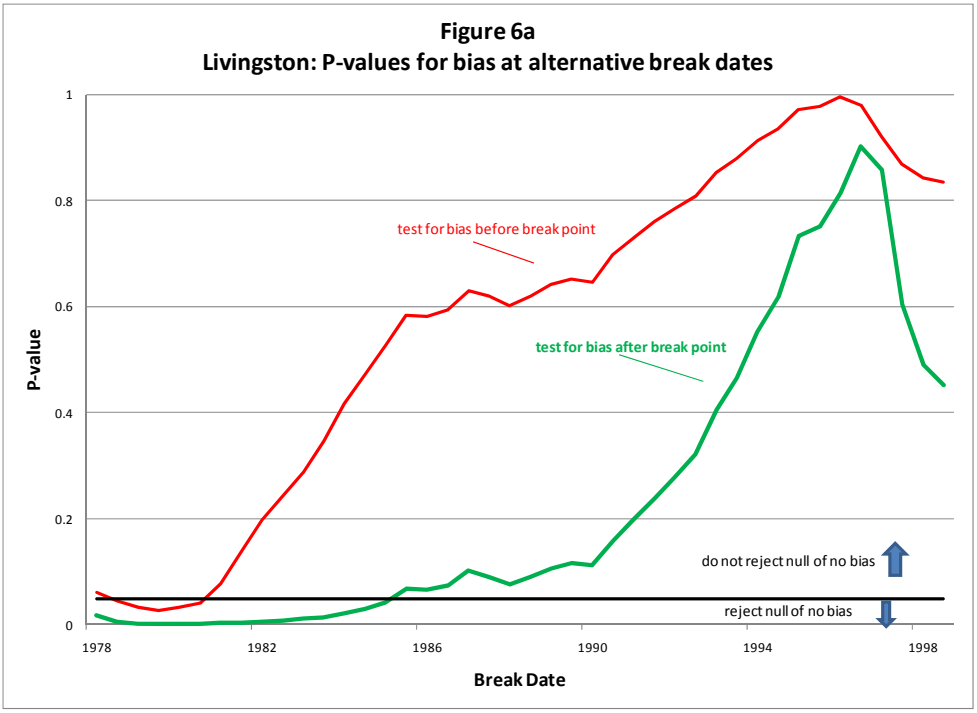
[9] Results from cutting the survey into non-overlapping observations can be found in the working-paper version, Croushore (2006).

Table 2
Test for Bias

Sample period	$\hat{\alpha}$	*p-value*	*N*	*reject null?*
Livingston Survey, Actuals = One-Quarter Later				
Full sample: 71H1-06H2	-0.014 (0.256)	.955	72	no
Survey of Professional Forecasters, Actuals = One-Quarter Later				
Full sample: 71Q1-06Q4	-0.007 (0.273)	.980	144	no

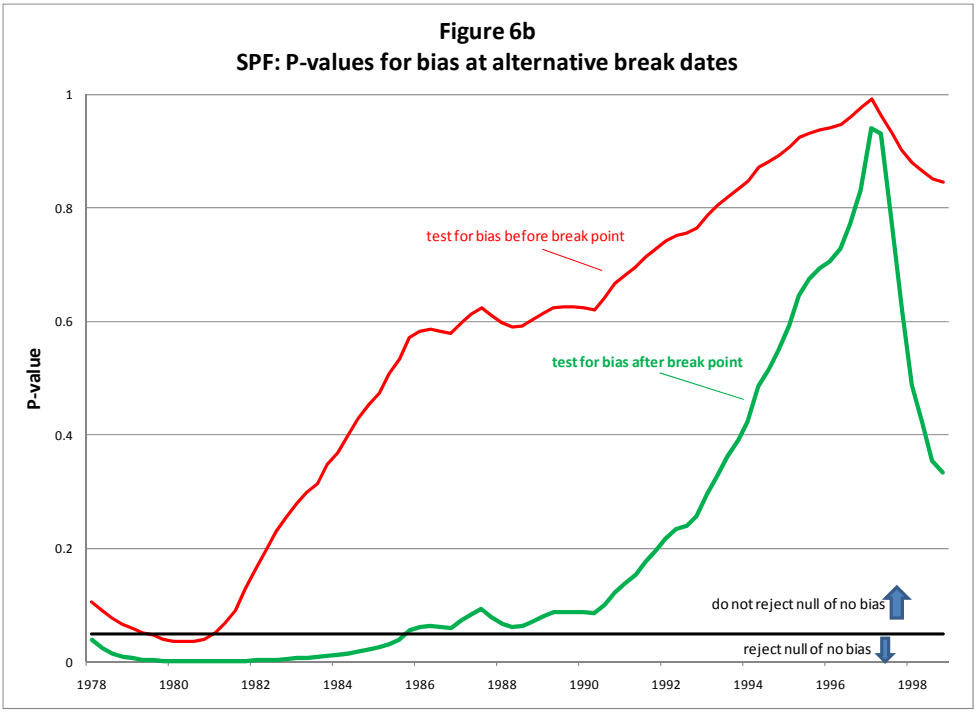
Note: standard errors are in parentheses. The null hypothesis is that there is no bias in the survey.

Figure 6, panel a shows the p-value of the test for the Livingston survey, while panel b shows results for the SPF. We test for bias in the sample period before the break, as well as the sample period after the break.



For the Livingston Survey, shown in Figure 6a, the line labeled "test for bias before break point" shows the p-value of the bias test for the sample from 1971:H1 to the date shown on the horizontal axis. The line labeled "test for bias after the break point" shows the p-value of the bias test for the sample from the date shown on the horizontal axis to 2006:H2. We reject the null hypothesis of unbiased forecasts only for samples that begin in 1971:H1 and end any time from 1978:H2 to 1980:H2; and for samples that end in 2006:H2 and begin any time from 1978:H1 to 1985:H1.[10] Clearly, the Livingston Survey shows bias only in the early years of the sample.

[10] The results reported here differ somewhat from the earlier literature, in part because our sample period for the SPF starts in 1971, while earlier papers began with samples in 1968. But from 1968 to 1971, the SPF deflator forecasts were rounded to the nearest whole digit, making the resultant inflation series quite volatile because of excessive rounding. The results of the earlier literature may have been overly influenced by that rounding effect. In addition, the Livingston survey in the early years employed a questionable methodology, in which the journalist in charge of the survey modified the forecasts when new data were released. Only after the Federal Reserve Bank of Philadelphia took over running the survey was the methodology made consistent. Thus early years of the Livingston survey are of questionable value.



For the SPF, shown in Figure 6b, we reject the null hypothesis of unbiased forecasts only for samples that begin in 1971:Q1 and end any time from 1979:Q3 to 1980:Q4; and for samples that end in 2006:Q4 and begin any time from 1978:Q1 to 1985:Q3. These dates are similar to those found for the Livingston survey. Again, most bias in the survey occurred in the early years.

The results here suggest that bias in the survey arose mainly in the 1970s and early 1980s (as suggested by Figures 1 and 2) and that the surveys have performed much better since then. But, this method takes for granted a sample that always begins in 1971 for the pre-break point and ends in 2006 for the post-break point. An alternative method would be to examine rolling windows of various sizes to see if there is evidence of bias in particular sections of rolling windows. For example, we could take 5-year rolling windows beginning with the sample from 1971:H1 to 1975:H4, then 1971:H2 to

1976:H1, and so on. Results of doing this exercise for both surveys for both 5-year and 10-year rolling windows are shown in Figures 7a and 7b.

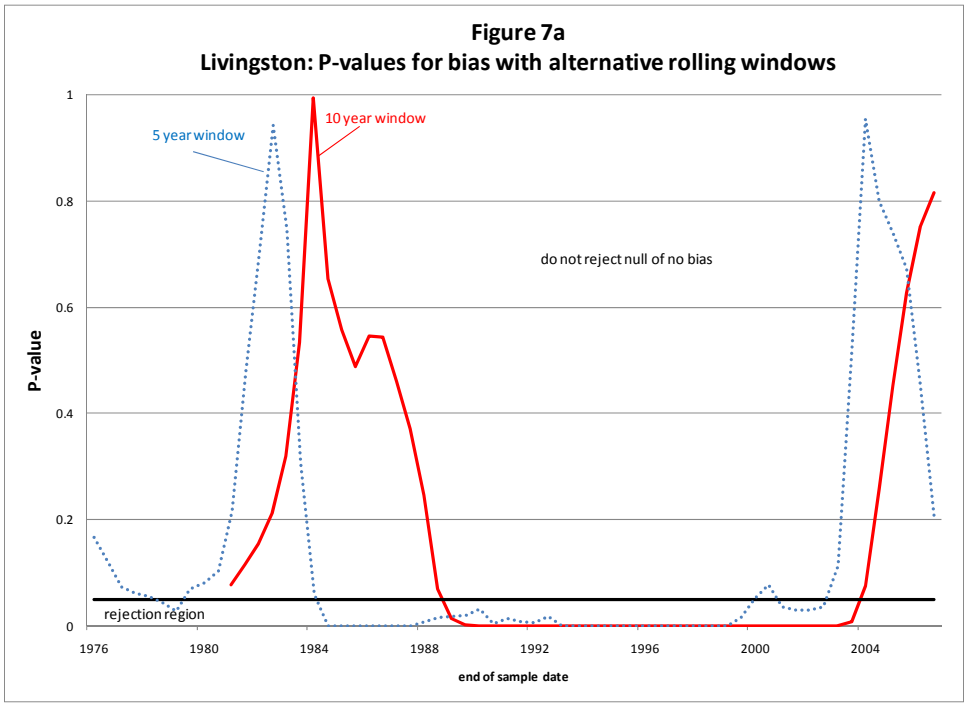
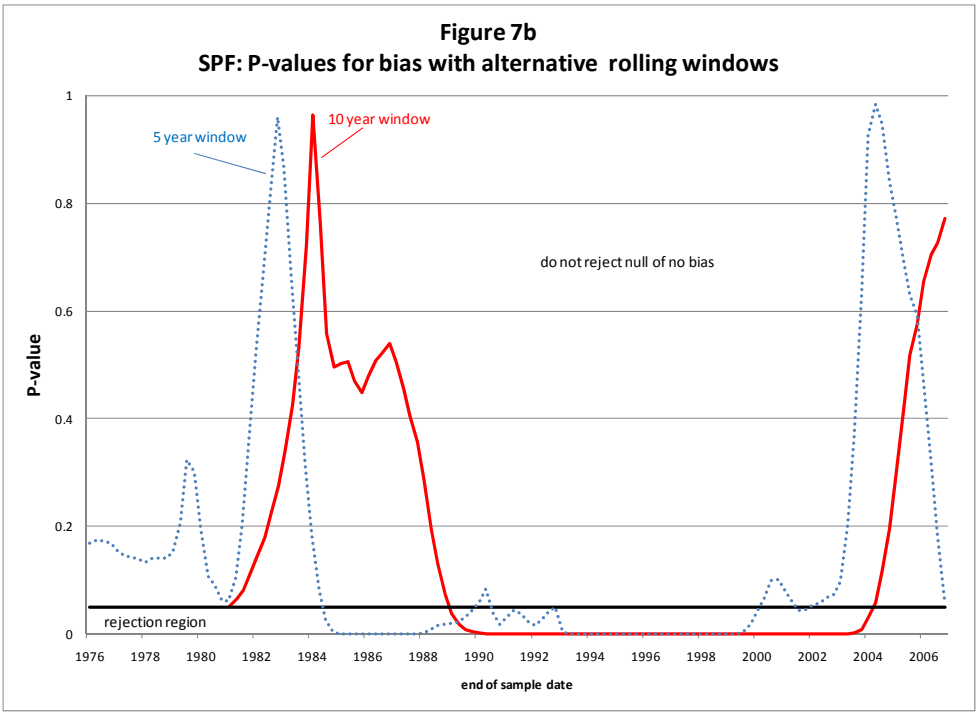


Figure 7a shows that in the very early part of the sample period the Livingston forecasts appeared biased, but that bias disappears in both the 5-year and 10-year windows in the early 1980s. More important and long-lasting biases appear later. They arise in the 5-year window beginning in 1984 and continuing through the end of the 1990s. The 10-year window identifies bias beginning in 1989 and continuing to 2003. Only in the 2000s does the bias finally disappear. Thus the statistics that we looked at earlier that included data through 2006 may have been misleading about the degree of bias in the survey.



Similar results occur for the SPF, as shown in Figure 7b, though the biases do not last as long as they did for the Livingston survey. Still, the 5-year window shows bias in the last half of the 1980s and for most of the 1990s, while the 10-year window shows bias beginning in 1989, continuing throughout the 1990s and ending only in 2004. These results suggest that an observer might be able to improve on the survey forecasts by testing for bias and then adjusting the forecasts appropriately.

FORECAST-IMPROVEMENT EXERCISES

The next question we seek to answer is: if you observed the pattern of past forecast errors, could you have used the knowledge to make better forecasts? Consider trying to improve on the forecasts in the following way. Run the bias regression in Equation (2), estimate $\hat{\alpha}$, and then create a new and improved forecast, π_t^i:[11]

$$\pi_t^i = \hat{\alpha} + \pi_t^f . \tag{3}$$

Those who argued in the early 1980s that the forecasts were irrational suggested that this approach would have led forecasters to have much smaller forecast errors than in fact they had. But suppose we had followed their advice over time. How big would the subsequent forecast errors be? And would following this advice lead to a lower root-mean-squared forecast error (RMSFE)? [12]

Running this experiment, using real-time data, leads to the results shown in Table 3. We try the forecast-improvement exercise in three different ways. The "full sample" results use all the forecasts starting in 1971 in a rolling fashion. We begin trying to improve the forecasts with the forecasts made in 1983:H1 for the Livingston Survey and 1978:Q1 for the SPF, to ensure that our regression has at least 20 observations. We run the regression given by Equation (2), use the estimate of $\hat{\alpha}$, then create a new and improved forecast, π_t^i according to Equation (3). Then we step forward one period (which

---

[11] This method of improving the forecast is based just on the bias term. You could alternatively estimate Equation (1), with both a bias estimate and a slope estimate. But in my experiments, this led to much worse forecasts than just using the bias term.

[12] It is also possible that the predictability of revisions to the data could lead to predictable forecast errors when using revised data, but not when using real-time data. There is some evidence that revisions are predictable in real time, as explored by Aruoba (2008) and Faust et al. (2005). We do not examine that potential connection in this paper.

is one half-year, so that the one-year-ahead forecasts will overlap), add one more data point to the sample, then rerun Equation (2) and form a new and improved forecast from Equation (3). For each date, we collect just the one-year-ahead forecast formed at that date. We proceed in this fashion through the end of the sample. Then, we calculate the root-mean-squared forecast error (RMSFE) of the forecasts made by the forecast-improvement method and compare them with the RMSFE of the survey itself. We test the statistical significance of the difference in RMSFEs using the Harvey et al. (1997) modification of the Diebold-Mariano (1995) procedure. The columns labeled "10-year Window" and "5-year Window" repeat the exercise described above. But instead of basing the forecast adjustment on the estimated bias term over the sample beginning in 1971, they use rolling windows of ten years and five years of forecasts, respectively. The idea here is that the degree of bias in the forecasts may have changed over time, as suggested by Figure 7, so this method allows us to ignore the older forecasts in attempting to improve later forecasts.

The results show that the use of Equation (3) to improve the survey forecasts is not very fruitful. The RMSFE for any attempt at forecast improvement is higher than the RMSFE for the original survey, though in no case is the RMSFE significantly higher.

Table 3
RMSFEs for Forecast-Improvement Exercises

			Attempts to Improve on Survey		
Survey	Period	Original Survey	Full Sample	10-year Window	5-year Window
Livingston	1983:H1–2006:H1	0.77	1.06 (0.11)	1.01 (0.29)	0.92 (0.26)
SPF	1978:Q1–2006:Q3	1.08	1.44 (0.19)	1.35 (0.35)	1.24 (0.39)

Note: numbers in parentheses below RMSFEs are p-values for the test of equal RMSFEs between the attempt to improve on the survey and the original survey, based on the Harvey-Leybourne-Newbold modification of the Diebold-Mariano procedure. Using one-quarter-later values as actuals.

The results are somewhat sensitive to the sample period chosen. However, if the sample begins earlier, there are fewer observations in the early part of the sample period, so the attempt at forecast improvement is thwarted by additional sampling uncertainty. If the sample begins later, the surveys are more accurate, and there is less to improve upon, as the period of substantial forecast errors in the survey drops out of the sample.

An alternative possibility would be to combine the information from Figure 7 and just attempt to improve on the survey when the rolling window suggests bias by having a p-value from the bias test of less than 0.05, but using the unadjusted survey forecast when the p-value in the rolling window is greater than 0.05. The results of doing so are shown in Table 4.

Table 4
RMSFEs for Forecast-Improvement Exercises
Modified to Adjust for Bias in Survey Forecasts If P-value < 0.05

			Attempts to Improve on Survey	
Survey	Period	Original Survey	10-year Window	5-year Window
Livingston	1983:H1–2006:H1	0.77	0.81 (0.08)	0.89 (0.05)
SPF	1978:Q1–2006:Q3	1.08	1.12 (0.05)	1.18 (0.08)

Note: numbers in parentheses below RMSFEs are p-values for the test of equal RMSFEs between the attempt to improve on the survey and the original survey, based on the Harvey-Leybourne-Newbold modification of the Diebold-Mariano procedure. Using one-quarter-later values as actuals.

The results in Table 4 show that only trying to improve the survey in the rolling windows when the p-value of the bias test is less than 0.05 is helpful in reducing the RMSFE. However, it is still not good enough to reduce the RMSFE below that of the survey. The difficulty is that even when the p-value is below 0.05, the bias may be changing, which is almost impossible to pick up in real time; the result is some particularly bad forecasts when the attempt to improve on the forecasts pushes them in the wrong direction.

PEARCE TEST

Perhaps the most convincing study of rational expectations in the 1970s was Pearce (1979). Pearce made the observation that, looking at CPI data, if you were to simply estimate an ARIMA model using standard Box-Jenkins techniques, you would create better forecasts than the Livingston Survey had. This test was so simple that it convinced even the most diehard supporter of rational expectations that something was wrong with the survey forecasts. But, again, Pearce's sample period happened to be fairly short and was one in which inflation generally rose unexpectedly. With more than 20 more years of data, is Pearce's result still valid? Ang et al. (2007) showed for the case of the CPI, survey forecasts were superior to various ARIMA models. We can run a similar exercise for the GDP inflation rate using real-time data.

Pearce assumed that the model appropriate for inflation was an IMA(1,1) process.[13] We run one set of experiments based on that process, and another in which we assume an AR process that is determined by calculating SIC values period by period, thus allowing the model to change over time.

The results of the exercise are shown in Table 5. Root mean squared forecast errors are higher for the ARIMA models than they are for the survey. There is little support for the view that a simple ARIMA model can do better than the survey forecasts. Using real-time data, no ARIMA model does better than the original survey, and some ARIMA models do significantly worse. Thus, there is no way that a forecaster in real

[13] See Stock and Watson (2007) for a detailed analysis of forecasting inflation using an IMA(1,1) process and other processes. Despite its simplicity, the IMA(1,1) model performs almost as well as other models developed after post-sample peeking, with the best model being an unobserved components/stochastic volatility model, which is similar to an IMA(1,1) model with a rolling window.

time could have used an ARIMA model to improve on the survey forecasts, over the entire sample period or from 1971 to 1981. Not only are surveys better than ARIMA models, they are better in real time, sometimes significantly so.

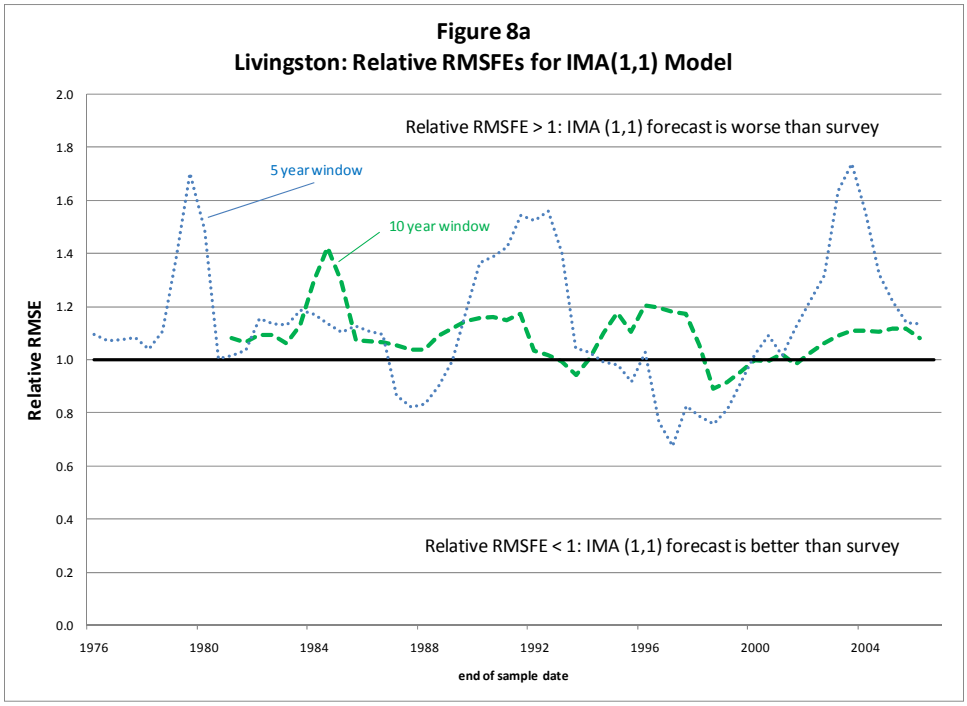
Table 5
Pearce Test
Root Mean Squared Forecast Errors

Sample period	Survey	IMA(1,1)	SIC
Survey of Professional Forecasters			
1971:Q1–1981:Q1	2.57	2.81 (0.57)	3.50 (0.00)
1971:Q1–2006:Q4	1.61	1.74 (0.51)	2.02 (0.09)
Livingston Survey			
1971:H1–1981:H1	2.17	2.35 (0.53)	3.27 (0.01)
1971:H1–2006:H1	1.39	1.51 (0.35)	1.95 (0.06)

Note: numbers in parentheses below RMSFEs are p-values for the test of equal RMSFEs between the ARIMA model and the original survey, based on the Harvey-Leybourne-Newbold modification of the Diebold-Mariano procedure. Actuals are one-quarter after the initial release.

Given the subsample differences found above in our bias tests, it may not surprise us to find that the Pearce test's results depend strongly on the sample. The test is fairly simple, and we can examine rolling 5-year windows and 10-year windows, to compare the root-mean-squared forecast errors of the survey forecast compared with the ARIMA model over 5-year and 10-year periods. We can also test for statistically significant differences in the RMSFEs. The results for the IMA(1,1) model compared

with the Livingston survey are shown in Figure 8a, which shows the relative RMSFE (the RMSFE of the IMA(1,1) model divided by the RMSFE of the Livingston survey) for 5-year and 10-year sample periods ending with the sample date shown on the horizontal axis. Again, we use the Harvey et al. modification of the Diebold-Mariano test for statistically significant differences in the root-mean-squared forecast error, shown in Figure 8b.



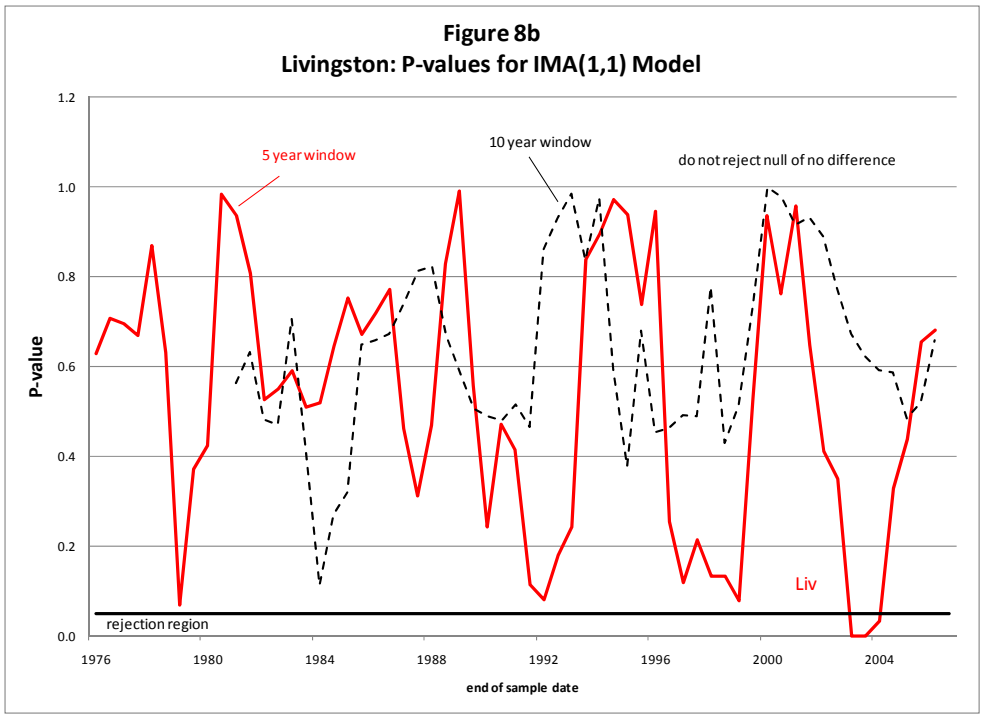
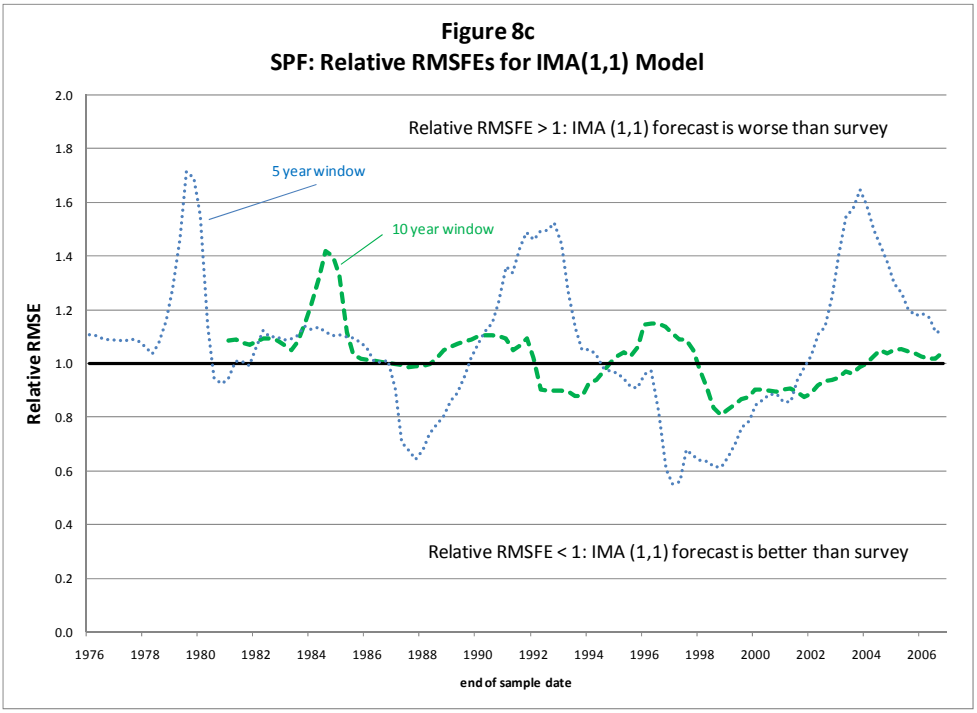
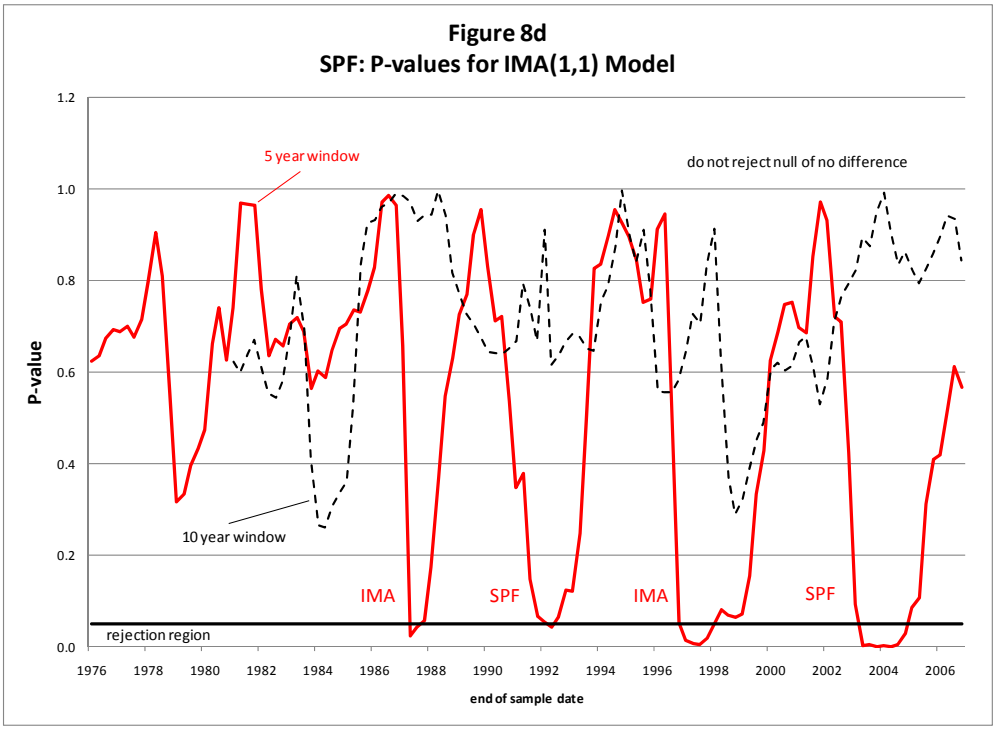


Figure 8a shows that most of the time for both 5-year and 10-year windows, the RMSFE of the survey is less than that of the IMA(1,1) model; in fact, that occurs 75% of the time for 5-year windows and 86% of the time for 10-year windows. Perhaps because the windows are fairly short, only a few of the differences are statistically significant (Figure 8b), showing that the Livingston survey has a significantly lower RMSFE than the IMA(1,1) model in the early 2000s, using a 5-year window of data.

Repeating this exercise with the SPF yields the results shown in Figures 8c and 8d. As was the case with the Livingston survey, most of the time the survey is better than the IMA(1,1) model; 64% of the time for 5-year windows and 61% of the time for 10-year windows. In this case, there are two periods in which the IMA(1,1) model does significantly better than the survey for 5-year windows (samples ending from 1987Q2 to 1987Q3 and 1996Q4 to 1998Q1); there are two periods in which the survey does

significantly better than the IMA(1,1) model (samples ending from 1992Q1 to 1992Q2 and 2003Q2 to 2004Q4).





We can perform the same types of exercises using the SIC to determine the appropriate size of AR model to forecast inflation. The results are similar to those of the IMA model, but are not shown to conserve space. For the Livingston survey, for most of the 5-year and 10-year rolling samples (85%), the survey does better than the SIC-chosen model, statistically significantly so in several cases. For the SPF, the results are a bit better for the SIC-chosen model, but still the SPF is better in about 80% of all the samples. Only for a few samples is the SIC-chosen model significantly better than the SPF for some 5-year windows; but the SPF is significantly better for many more samples.

CONCLUSION

The Livingston Survey and the Survey of Professional Forecasters developed poor reputations because of the systematic pattern of forecast errors found in the 1970s. Using basic statistical tests, researchers found that the forecast errors from the surveys failed to pass a number of basic tests, most importantly the bias test and the Pearce test. But when we look at a much longer sample of data, which goes beyond the years in which movements of inflation were dominated by oil-price shocks and bad monetary policy, we find that the inflation forecasts pass those statistical tests. However, for some subsamples, the surveys do show evidence of persistent errors. Nonetheless, even attempting to account for the short-run bias in the survey forecasts, we were not able to do better than the survey forecasts in terms of reducing root-mean-squared forecast error.

Why might we observe bias in the survey forecasts for certain sub-samples, but not enough bias to make it possible to improve on the forecasts in real time? The most likely explanation is that structural breaks have occurred that the forecasters learned

about only gradually. In that scenario, the forecasters could have rational expectations, but as they learn about the break their forecasts exhibit persistent errors of the same sign. This possibility has been explored in the literature by Evans-Wachtel (1993) in the context of a Markov switching model of inflation regimes, and Orphanides-Williams (2005), who develop a learning model by which inflation expectations are formed. Either mechanism would lead to persistent short-run forecasting errors in response to a structural shock. For example, in second half of the 1990s, forecasters in the surveys did not adjust their forecasts of potential output up until 1999, even though in retrospect potential output growth increased beginning in 1996. This slow learning may have led them to believe that output was outstripping potential output, leading them to forecast higher inflation, when in fact inflation was declining. But eventually they caught on and raised their forecasts of potential output, and the forecast errors returned to near zero.

The results in this paper complement the results of Ang et al. (2007), who found that survey forecasts are generally superior to forecasts formed by other methods, including ARIMA models, models based on the Phillips curve, and term-structure models. Their results were convincing for CPI-based forecasts, though they did not use real-time data in evaluating forecasts using the GDP inflation rate, unlike this paper. Combining the Ang et al. results with those in this paper, one could conclude that the negative view that economists had of survey forecasts, based on the literature in the early 1980s, was unwarranted.

Why might survey forecasts outperform other models overall? Forecasters clearly use statistical models in making their forecasts, and then judgmentally adjust the forecasts. Fildes-Stekler (2002) suggest four reasons for the judgmental adjustment: (1)

dealing with structural breaks; (2) handling changes in parameters; (3) dealing with data revisions; and (4) using more information that is not captured in the statistical model. The literature investigating the role of judgment in adjusting models suggests that such judgment is important in reducing forecast errors; see, for example, Wallis (1989) and Zarnowitz (1992).

This is not to say that there are no issues with forecast bias and inefficiency, just that the apparent bias found in the early literature has dissipated. There is much room for evaluating these same forecasts with respect to their efficiency related to changes in monetary policy (see Ball-Croushore (2003), for example), the orthogonality of forecast errors to other variables known at the time the forecasts were made, and the possible bias in particular sub-samples of the data. Deeper investigation of these issues is certainly warranted.

REFERENCES

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.

Aruoba, S. Boragan. "Data Revisions Are Not Well Behaved." *Journal of Money, Credit, and Banking* 40 (March-April 2008).

Baghestani, Hamid M., and Amin M. Kianian. "On the Rationality of US Macroeconomic Forecasts: Evidence from a Panel of Professional Forecasters." *Applied Economics* 25 (1993), pp. 869–878.

Ball, Laurence, and Dean Croushore. "Expectations and the Effects of Monetary Policy." *Journal of Money, Credit, and Banking* 35 (August 2003), pp. 473–484.

Boskin, Michael, et al. *Toward a More Accurate Measure of the Cost of Living* (1996), U.S. Senate Finance Committee.

Brown, Bryan W., and Shlomo Maital. "What Do Economists Know? An Empirical Study of Experts' Expectations," *Econometrica* 49 (March 1981), pp. 491-504.

Bryan, Michael F., and Stephen G. Cecchetti, "Measuring Core Inflation," in N. Gregory Mankiw, ed., *Monetary Policy*. Chicago: University of Chicago Press, 1994, pp. 195-215.

Capistrán, Carlos, and Allan Timmermann. "Disagreement and Bias in Inflation Expectations." Working paper, Bank of Mexico, June 2006.

Capistrán, Carlos, and Allan Timmermann. "Forecast Combination with Entry and Exit of Experts." Working paper, Bank of Mexico, April 2007.

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. 3-15.

Croushore, Dean. "The Livingston Survey: Still Useful After All These Years," Federal Reserve Bank of Philadelphia *Business Review*, March/April 1997, pp. 15-27.

Croushore, Dean. "Forecasting with Real-Time Macroeconomic Data." In: Graham Elliott, Clive W.J. Granger, and Allan Timmermann, eds., *Handbook of Economic Forecasting* (Amsterdam: North-Holland, 2006).

Croushore, Dean. "An Evaluation of Inflation Forecasts from Surveys Using Real-Time Data." Federal Reserve Bank of Philadelphia Working Paper No. 06-19, October 2006.

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

Croushore, Dean, and Tom Stark, "A Real-Time Data Set for Macroeconomists: Does the Data Vintage Matter?" *Review of Economics and Statistics* 85 (August 2003), pp. 605–617.

Davies, Antony. "A Framework for Decomposing Shocks and Measuring Volatilities Derived from Multi-Dimensional Panel Data of Survey Forecasts," *International Journal of Forecasting* 22 (April/June 2006), pp. 373–393.

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

Dua, Pami, and Subash C. Ray. "ARIMA Models of the Price Level: An Assessment of the Multilevel Adaptive Learning Process in the USA." *Journal of Forecasting* 11 (September 1992), pp. 507–516.

Evans, Martin, and Paul Wachtel. "Inflation Regimes and the Sources of Inflation Uncertainty." *Journal of Money, Credit and Banking* 25 (August 1993, part 2), pp. 475–511.

Faust, Jon, John H. Rogers, and Jonathan H. Wright. "News and Noise in G-7 GDP Announcements." *Journal of Money, Credit, and Banking* 37 (June 2005), pp. 403-419.

Fildes, Robert, and Herman Stekler. "The State of Macroeconomic Forecasting." *Journal of Macroeconomics* 24 (November 2002), pp. 435-468.

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.

Harvey, David, Stephen Leybourne, and Paul Newbold. "Testing the Equality of Prediction Mean Squared Errors." *International Journal of Forecasting* 13 (June 1997), 281-291.

Keane, Michael P., and David E. Runkle. "Testing the Rationality of Price Forecasts: New Evidence From Panel Data," *American Economic Review* 80 (1990), pp. 714-35.

Maddala, G.S. "Survey Data on Expectations: What Have We Learnt?" in Marc Nerlove, ed., *Issues in Contemporary Economics, vol. II. Aspects of Macroeconomics and Econometrics.* New York: New York University Press, 1991.

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.

Orphanides, Athanasios, and John C. Williams. "Imperfect Knowledge, Inflation Expectations, and Monetary Policy." In *Inflation Targeting*, ed. by Ben S. Bernanke and Michael Woodford (Chicago: University of Chicago Press, 2005), pp. 201-234.

Pearce, Douglas K. "Comparing Survey and Rational Measures of Expected Inflation," *Journal of Money, Credit and Banking* 11 (November 1979), pp. 447-56.

Rhim, Jong C., Mohammed F. Khayum, and Timothy J. Schibik. "Composite Forecasts of Inflation: An Improvement in Forecasting Performance." *Journal of Economics and Finance* 18 (Fall 1994), pp. 275–286.

Stock, James H., and Mark W. Watson. "Why Has U.S. Inflation Become Harder to Forecast?" *Journal of Money, Credit, and Banking* 39 (February 2007), pp. 3–34.

Theil, Henri. *Applied Economic Forecasting*. Amsterdam: North Holland, 1966.

Thomas, Lloyd B., Jr. "Survey Measures of Expected U.S. Inflation," *Journal of Economic Perspectives* 13 (Fall 1999), pp. 125–144.

Vanderhoff, James. "A 'Rational' Explanation for 'Irrational' Forecasts of Inflation." *Journal of Monetary Economics* 13 (May 1984), pp. 387–392.

Wallis, Kenneth F. "Macroeconomic Forecasting: A Survey." *Economic Journal* 99 (March 1989), pp. 28–61.

Webb, Roy H. "The Irrelevance of Tests for Bias in Series of Macroeconomic Forecasts," Federal Reserve Bank of Richmond *Economic Review* (November/December 1987), pp. 3-9.

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

Zarnowitz, Victor. *Business Cycles: Theory, History, Indicators and Forecasting*. National Bureau of Economic Research Studies in Business Cycles, vol. 27 (Chicago: University of Chicago Press, 1992).