

An Evaluation of Recent Macroeconomic Forecast Errors

Despite a significant decline in the pace of economic growth in the second half of 2000, macroeconomic forecasters underpredicted real GDP growth and overpredicted the unemployment rate by a significant amount, for the fifth consecutive year. On average, real GDP forecasts were about 2 percentage points below the actual data for the 1996–2000 period, and unemployment rate forecasts about 0.5 percentage point above. On a more positive note, forecasters ended their chronic overprediction of inflation during much of this period. Nevertheless, surprisingly large and persistent errors in recent forecasts of GDP, inflation, and unemployment have perplexed macroeconomists and policymakers for quite some time, and they merit closer examination.¹

To begin, we ask whether the large, persistent, and one-sided errors observed recently violate the principal objectives of economic forecasting in a statistically significant way. And if they do, what caused these significant errors and are they likely to continue? Such violations undercut the credibility of forecasting models and complicate the already difficult task of setting appropriate monetary policy. Efforts to discern the causes of recent errors and to improve macroeconomic forecasting models may help improve the conduct of policy.

This article evaluates recent forecast errors made by private forecasters in an attempt to understand why forecasts have gone so far awry. The investigation centers on errors in forecasts of real GDP growth, inflation, the unemployment rate, and nominal and real short-term interest rates since 1969. The focus is on one-year-ahead forecasts because lags in the effects of monetary policy require the Federal Reserve to forecast economic activity well in advance when setting its current interest rate target. Average forecasts, which tend to yield smaller errors, and individual forecasts are examined.

Scott Schuh

Senior Economist, Federal Reserve Bank of Boston. The author thanks Jennifer Young for excellent research assistance, Delia Sawhney and Georgeanne DaCosta for assistance in collecting and compiling data, and Lynn Browne, Jeff Fuhrer, Simon Gilchrist, Richard Kopcke, Stephen McNees, and Geoff Tootell for helpful comments.

A primary motivation for looking at the performances of individual forecasters is to learn whether some forecasters are better than others and, if so, what makes them better. In particular, did any forecasters *not* make large errors in the late 1990s and, if so, should we pay more attention to their forecasts for 2001? This endeavor extends the work of Zarnowitz (1985), McNees (1992), Zarnowitz and Braun (1993), and others who have used forecast data from panels of forecasters (Survey of Professional Forecasters and *The Wall Street Journal*) to assess relative individual forecaster performance. This study updates both data sources through 2000.

The strategy is to conduct standard diagnostic tests of average and individual macroeconomic forecasts in search of evidence of statistically significant problems that would explain recent forecast errors. One test determines whether forecasts are unbiased, meaning that they tend to be consistently neither too high nor too low relative to the actual data. Another test determines whether forecasts are efficient, meaning that no additional information was readily available to forecasters that could have been used to make forecasts more accurate. If forecasts are not efficient, forecast errors will exhibit certain types of correlation that could be exploited to improve forecasts. One or both of these tests should detect breakdowns or ongoing shortcomings in the models used by forecasters.

*Lags in the effects of monetary policy
require the Federal Reserve
to forecast economic activity well
in advance when setting its current
interest rate target.*

Although recent forecast errors have been large and troubling, average macroeconomic forecasts generally have been *unbiased* during the past three decades. This finding contrasts with most previous studies of average forecasts, apparently because the

¹ In his May 6, 1998, speech at the Chicago Fed international banking conference, Federal Reserve Chairman Alan Greenspan said: "Forecasts of inflation and of growth in real activity for the United States, including those of the Federal Open Market Committee, have been generally off for several years. Inflation has been chronically overpredicted and real GDP growth underpredicted."

sample period is much longer, so that large one-sided errors over shorter periods, such as the recent underpredictions of GDP, average out over time. Some sample periods beginning in the early 1980s show modest evidence of statistically significant bias in recent GDP forecasts. But the evidence is not robust across sample periods, and it does not appear in other macroeconomic variables.

However, there is ample evidence that average macroeconomic forecasts are not efficient. Standard regression-based tests indicate that most forecasts—with the notable exception of real interest rate forecasts—do not properly or completely incorporate readily available information on current macroeconomic forecasts or past macroeconomic data. These tests cannot pinpoint why forecasts exhibit this problem without detailed descriptions of the underlying forecasting models. But the tests do reveal that adding this information to macroeconomic forecasts significantly reduces the magnitude and volatility of historical forecast errors and may explain the recent behavior of the errors.

Indeed, half or more of the average GDP forecast error from 1996 to 2000 can be attributed to inefficiency in average GDP forecasts. Specifically, the evidence shows that forecasters tend to underpredict GDP significantly when inflation and nominal interest rates are unusually low, and vice versa. Adjusting recent GDP forecasts for inflation and interest rates reduces the 1.9 percent average forecast error by about one-half to two-thirds (0.9 to 1.2 percentage points).² This adjustment is large because inflation and interest rates were unusually low, and inflation forecast errors unusually large, during this period. Because the inefficiency of GDP forecasts existed prior to 1996, a significant portion of the average GDP error did *not* result from a sudden breakdown of GDP forecasts in the late 1990s.

The hefty remainder of the average GDP error—up to 1 percentage point—is not explained readily by other basic macroeconomic information available to forecasters. This remainder may have resulted from some sort of breakdown in recent GDP forecasts. One reasonable possibility is that trend productivity growth increased by approximately this amount but the increase was unknown to forecasters and excluded from their GDP forecasts. Indeed, by 2000 many forecasters explicitly began revising their trend productiv-

² Unfortunately, correcting for inefficiencies does not systematically improve inflation forecasts during this period, although it does cut in half the variance of inflation forecast errors by reducing very large errors in the 1970s and early 1980s.

ity estimates upward, and these revisions were reflected in higher GDP forecasts.

Individual forecasters fared no better, in general, than the average of all forecasters. Between 1996 and 2000, essentially all individual forecasters underpredicted GDP while nearly all forecasters overpredicted unemployment, and essentially all forecasters overpredicted inflation in 1997 and 1998. Because no one got it “right,” individual forecasters offer little hope of learning what caused the unusually large errors.

Nevertheless, individual forecasters exhibit distinct differences in their ability to produce unbiased and efficient forecasts. Regression-based tests show that most individual forecasts of GDP and inflation tend to be either unbiased and efficient—“good” forecasts—or biased and inefficient—“bad” forecasts. However, the time period of the forecast seems to be a key determinant of individual performance. Overall, this preliminary evidence motivates further investigation of individual forecasts.

I. Basic Principles of Economic Forecasting

Before analyzing the data, it is helpful to review briefly some of the basic principles of forecasting using econometric models.³ Traditionally, economists have postulated that the central goal is to produce unbiased and efficient forecasts with uncorrelated forecast errors. These properties stem from the assumption that forecasters use all available information and use it correctly. If so, errors made by econometric forecasting models will exhibit certain statistical properties that can be measured and observed. Forecasts that exhibit these properties suggest that the forecasters indeed use accurate models and all available information when making forecasts.

Not all forecasters necessarily share this traditional forecasting goal because they may have other objectives. Laster, Bennett, and Geoum (1999) find evidence that publicity may be more important than accuracy to some forecasters, presumably because it might attract clients and increase profits. Lim (2001) finds evidence that financial analysts bias upward their forecasts of corporate earnings in exchange for inside information that improves the accuracy of their forecasts. Lamont (1995) finds evidence that reputation effects lead forecasters to produce more radical, less accurate forecasts as they grow older and more established.⁴ Some forecasters may place a higher weight on predicting recessions than rapidly expanding activity, and still others may shade their forecasts toward the consensus to

avoid making unusual errors. Even the monetary authority might have nontraditional forecasting objectives—perhaps it would accept biased inflation forecasts in exchange for more accurate forecasts of rising inflation—but such nontraditional theories have not been widely developed yet.

Let y_t be a variable, such as output growth, to be forecast. The forecasting model for forecaster i is

$$y_t = f_i(A_i(L)y_{t-1}, B_i(L)X_{it}; \delta) + \epsilon_{it}, \quad (1)$$

where A_i and B_i are unknown parameters, L is the lag operator; X_{it} is a set of explanatory variables, δ is a set of parameters, $f_i(\cdot)$ is a function (possibly nonlinear), t denotes time, and ϵ_{it} is a stochastic error.⁵ Every element of equation (1) has a subscript i to indicate that all aspects of the forecasting model can differ across forecasters.⁶

Forecasts are obtained as follows. Using data on y_t and X_{it} , forecasters calculate econometric estimates of δ (note that estimation techniques also may vary across forecasters). Using a variety of methods, forecasters obtain forecasts (denoted by a tilde) of the explanatory variables, $\tilde{X}_{i,t+k}$, k periods into the future.⁷ Then the k -step ahead forecast (made at the beginning of period t) is

$$\tilde{y}_{i,t+k} = f_i(A_i(L)y_{t+k-1}, B_i(L)\tilde{X}_{i,t+k}; \hat{\delta}_i) \quad (2)$$

and the k -step ahead forecast error is

$$\tilde{\epsilon}_{i,t+k} = y_{i,t+k} - \tilde{y}_{i,t+k}. \quad (3)$$

Under this definition, positive forecast errors indicate underprediction and negative errors indicate overprediction.

Differences among forecasters and forecasting

³ Pindyck and Rubinfeld (1976) and Granger and Newbold (1986) are good references with more details.

⁴ Lamont's results are obtained from a panel of forecasters published in *BusinessWeek* magazine but Stark (1997) does not replicate Lamont's result with the anonymous Survey of Professional Forecasters (SPF). The difference may be explained by the fact that reputation effects do not appear in confidential forecasts.

⁵ All of the ideas in this discussion generalize to multivariate forecasts as well. In fact, most economic forecasts come from multivariate models that produce simultaneous forecasts of many variables.

⁶ Even y_t could be viewed as varying across forecasters. For example, output could be real GDP or industrial production, the data could be levels or growth rates, and growth rates could be based on annual averages or four-quarter changes.

⁷ These methods include auxiliary forecasting models, judgmental extrapolation, exogenous policy assumptions, and forecasts from other forecasters.

models imply that individual forecasts and forecast errors will be different as well. The following basic principles of economic forecasting are used to evaluate the performance of forecasts:

- **Unbiased Forecasts**—Forecasts should be approximately equal to the actual data on average over time, $(1/T)\sum_{t=1}^T (y_t - \hat{y}_{it}) = 0$; thus forecast errors should be approximately zero on average over time, $(1/T)\sum_{t=1}^T \tilde{\epsilon}_{it} = 0$.
- **Efficient Forecasts**—Forecasts should come from accurate models of economic behavior that use all relevant information readily available to the forecaster. No other model or readily available information should be able to improve the forecasts.
- **Uncorrelated Errors**—Forecast errors should not be correlated with past errors ($\text{corr}(\epsilon_{it}, \epsilon_{i,t-s}) = 0$ for all periods $s > 0$) or with other information readily available to the forecaster.

Forecasters can be evaluated according to how well they adhere to these principles.

Violations of these basic forecasting principles can occur for many reasons. Inaccurate parameter estimates, omitted explanatory variables, and erroneous models (for example, linear instead of nonlinear) all can produce either bias or inefficiency or both in forecasts. Inefficient forecasts will produce correlated errors, but inaccurate parameter estimates or the inclusion of explanatory variables with spurious correlation can, too. Biased forecasts may or may not be inefficient, and vice versa. In short, these diagnostic tests only detect problematic forecasts but they cannot pinpoint the exact reason for the problem. One needs precise details on the forecasting model to produce more specific diagnoses.

One common statistic used to track forecast errors and identify bias or breakdowns in forecasting models is cumulative forecast errors. If forecasts become biased or a forecasting model breaks down, forecast errors will not average to zero any more and the sum of all errors will become large in absolute value. If the cumulative sum of forecast errors exceeds a statistically determined critical value, the forecasts are biased (this is called a CUSUM test). A preliminary indication of bias is that forecasts tend to become one-sided, as with the recent consecutive string of positive GDP errors.

⁸ Note that this test indirectly detects whether the forecast has failed to include the information contained in Z_{it} . It does not detect whether the forecasting model is incorrect or the parameter estimates are inaccurate, two conditions that might be classified as inefficiency as well.

Another common method of testing for bias and efficiency is a regression-based approach. The estimating model is

$$y_{t+k} = \beta_{0i} + \beta_{1i}\tilde{y}_{i,t+k} + \beta_{2i}Z_{it} + \eta_{i,t+k}, \quad (4)$$

where Z_{it} is a set of variables that were available when the forecast was made and could improve the forecast. The logic of this test is that the forecast should track the actual data one-for-one over time, hence $\beta_{1i} = 1$, and the other explanatory variables should not matter, hence $\beta_{0i} = \beta_{2i} = 0$. Assuming $\beta_{2i} = 0$, if $\beta_{0i} \neq 0$ there is evidence of bias. And if $\beta_{2i} \neq 0$ there is evidence of inefficiency. Formally, the two-step procedure is as follows. To test for bias, set $\beta_{2i} = 0$ and evaluate the null hypothesis $H_0: [\beta_{0i} \beta_{1i}] = [0 \ 1]$; then, to test jointly for bias and efficiency, evaluate the null hypothesis $H_0: [\beta_{0i} \beta_{1i} \beta_{2i}] = [0 \ 1 \ 0]$.⁸ The list of potential candidates for Z_{it} is extremely long and it is impossible to test all candidates. Consequently, in this study, the list is limited to the latest available data ($y_{i,t-1}$ and $X_{i,t-1}$) and other forecasts made at the same time.⁹

Past applications of this regression-based testing procedure often have led to rejections of unbiasedness, efficiency, or both for many macroeconomic variables. For examples, see Zarnowitz (1985), Baghestani and Kianian (1993), and Loungani (2000). But the rejections are not universal. A prominent exception is Keane and Runkle (1990), who found that price forecasts are unbiased and efficient (controlling for money, oil prices, and lagged prices). Bonham and Cohen (1995) showed that the Keane and Runkle efficiency tests were flawed and that price forecasts were inefficient. Keane and Runkle (1995) agreed but argued that the Bonham-Cohen methodology may have difficulties too. In short, the question of whether forecasts are unbiased and efficient remains open.

Keane and Runkle (1990) also raised a potentially important methodological point about the use of average forecasts in these regression-based tests. Regression tests that use average forecast data yield estimates of the β parameters that may contain aggregation bias, which is the difference between these estimates and the average of all parameter estimates obtained from regression tests that use individual forecast data. If this aggregation bias is large, it could prevent tests with average forecast data from providing accurate inference about the collective bias and efficiency of the

⁹ Another feasible candidate for the efficiency test is the dispersion of forecasts across individual forecasters, or uncertainty, as described in Zarnowitz and Lambros (1987).

group of forecasters (but not individual forecasters).¹⁰ Thus, the results in Section III of this article should be interpreted with caution.

To deal with the aggregation problem, Keane and Runkle (1990) propose a panel regression approach using cross-section time series data on individual forecasters. This methodology overcomes the shortcomings of panel regressions noted by Zarnowitz (1985) and its β parameter estimates do not exhibit aggregation bias. However, this approach has drawbacks, too. In particular, it eliminates aggregation bias by restricting slope parameters (β_{1i}, β_{2i}) to be the same for all forecasters. But this strong restriction is not satisfied in the data (see the appendix tables in Keane and Runkle 1990, for example) and it is the only case in which aggregation bias is certain to be eliminated, as discussed in Theil (1971). Panel regressions that impose counterfactual parameter restrictions may exhibit statistical problems as well.¹¹

For these reasons, and brevity, I do not explore the panel regression methodology. Nevertheless, my general conclusions from the average forecast data are exactly the same as those drawn from the Keane and Runkle and Bonham and Cohen work—forecasts are unbiased but inefficient. In addition, the continued use of average forecasts in the literature suggests a lack of consensus on this issue. Ultimately, the primary interest of this investigation is in determining whether some forecasters are better than others, and that involves estimating individual regression models for each forecaster separately.

II. Data Sources

The forecast data come from three sources, which are described in more detail in the Data Appendix. One is the Survey of Professional Forecasters (SPF),

¹⁰ Baghestani and Kianian (1993) argue that one can use median, rather than average, forecasts to circumvent this criticism. However, the median also depends on the distribution of individual forecasts, albeit in a more complicated manner, and thus does not avert the problem. I find that the test results are essentially the same when estimated with average and median forecasts.

¹¹ There may be other problems with the Keane-Runkle panel regression. First, it also makes strong assumptions about individual forecast errors that imply there are no performance differences across forecasters, which Section IV of this article shows may be counterfactual. Second, the individual forecaster data set forms an unbalanced panel because forecasters are not in the panel for all years. Thus, if forecasters in different time periods have different parameters, the panel estimates could be misleading, which Section IV of this article also shows may be a problem. Third, GMM estimation methods have been shown to exhibit serious biases in small samples. All of these issues require extensive additional investigation.

which begins in 1969 and includes approximately 40 anonymous forecasters per survey. A second source is *The Wall Street Journal* (WSJ). McNees (1992) originally developed this panel containing about 30 of the publicly identified forecasters with data beginning in the mid 1980s. The third source is the Blue Chip Economic Indicators (BC), which begins in 1977 and includes about 50 institutional forecasters whose identities also are revealed. Individual BC forecasts have not been compiled, so only the BC consensus (average) data are used here for comparison with the average SPF and average WSJ data. The composition of forecasters changes over time in each data source; management of the SPF changed as well.

One-year-ahead forecasts have received less attention than one-quarter-ahead forecasts but are at least as important, and possibly more so.

Forecasters differ in potentially important ways. Because SPF forecasters are anonymous and report the same forecasts they sell in the market, typically it is assumed that financial incentives make the SPF forecasts likely to be the most accurate of private forecasts (see Keane and Runkle 1990; Baghestani and Kianian 1993). In contrast, forecasters in the WSJ and BC surveys are identified by name and thus may have other objectives. Publicity is a motivating factor for BC forecasters, according to the Laster, Bennett, and Geoum (1999) study, and the WSJ survey emphasizes publicity by picking and publicizing the best forecasters each year (but not on average across years). Also, identification might lead some forecasters not to provide their best forecasts because they could lose income from potential clients. Yet another difference is that the WSJ forecasts come from individuals, whereas the BC forecasts come from institutions that may have employed different forecasters over time; the type of SPF forecaster is unknown.

This study focuses on forecasts of five key macroeconomic variables: real output growth, \tilde{y}_t ; inflation, $\tilde{\pi}_t$; the unemployment rate, \tilde{u}_t ; the nominal short-term interest rate, \tilde{r}_t , and the real short-term interest rate,

$\tilde{r}_t = \tilde{i}_t - \tilde{\pi}_t$.¹² The real rate is not reported explicitly by forecasters but is implied by their nominal rate and inflation forecasts. Real output is GNP prior to the early 1990s and GDP thereafter. The unemployment rate is for the total civilian labor force. Inflation is growth in the total CPI-U since the early 1980s and growth in the GNP implicit price deflator prior to that. The nominal short-term interest rate is mainly the 3-month Treasury bill rate. Each forecast is for one calendar year, made at the end of one year and covering the subsequent calendar year.¹³

One-year-ahead forecasts have received less attention than one-quarter-ahead forecasts but are at least as important, and possibly more so. Because changes in monetary policy typically affect the economy with a lag of six to 18 months, policymakers must evaluate economic activity in the future to determine the appropriate monetary conditions today. One-year-ahead forecasts are potentially more difficult than one-quarter-ahead forecasts because the opportunity for unexpected economic developments is greater. However, McNees (1992) reports mixed evidence on the precision of short-term versus medium-term forecast errors across a range of variables and individual forecasters.

Two types of actual data are used to calculate forecast errors. One is the most recent version of the data, which includes all revisions since the forecast period. These current data contain information unavailable to forecasters at the time of their forecasts—such as late-reported data or methodological changes—that could induce biases, inefficiencies, and correlation in forecast errors. The other type is so-called real-time data, which are those available when the forecasts were made. These data were obtained from the Federal Reserve Bank of Philadelphia and various issues of the *Economic Report of the President*.

Many analysts have argued that real-time data are most appropriate for evaluating forecast performance and monetary policy decisions.¹⁴ The argument is that forecasters should be expected to forecast economic

activity only as it was understood from (potentially erroneous) data at the time, but not data revisions or “true” economic activity. On the other hand, the optimal conduct of monetary policy is likely to depend on “true” economic activity, so it seems reasonable to examine forecasts relative to the best estimates of activity—presumably, the current data.

It turns out, however, that the dichotomy between current and real-time data is largely irrelevant, for two reasons. First, current and real-time data are virtually

The dichotomy between current and real-time data is largely irrelevant because they are only modestly different at the annual frequency.

the same for inflation, unemployment, and interest rates because these data are subject to very little revision. Second, current and real-time real GDP data are only modestly different at the annual frequency. Apparently, temporal aggregation smoothes many of the differences between these versions at quarterly frequencies. Actual data in this study are current data except in a few instances where real-time GDP data are noted (y_t^r).

III. Average Forecasts

To begin, it is helpful to examine the historical evidence on forecasts and errors by looking at data averaged across forecasters (sometimes called consensus forecasts). Figure 1 plots data, forecasts, and forecast errors for five macroeconomic variables over the past three decades. The left column shows the SPF forecasts and actual data; the right column shows the SPF forecast errors, plus the WSJ and BC errors for comparison. Table 1 reports basic statistics for the errors during the full sample and latest five years.

Despite the potential for substantive differences among the three forecast groups, it is apparent from the errors shown in Figure 1 that the three average forecasts are similar during their common sample periods. The WSJ inflation errors are slightly more variable than the other two inflation errors and the SPF interest rate error in 1982 is quite different, but

¹² If the real interest rate forecast were defined as $\tilde{r}_t = \tilde{i}_{t-1} - \tilde{\pi}_t$, as in some modern macroeconomic models, the inflation and real rate forecasts would be essentially the same. Moreover, this specification assumes the nominal rate is fixed but its maturity is only three months whereas the real rate forecast is for one year.

¹³ Although one-year-ahead forecasts can be constructed at higher frequencies, the overlapping nature of such forecasts makes them dependent over time and this correlation introduces potential problems for statistical and econometric analysis.

¹⁴ See, for examples, Keane and Runkle (1990), Robertson and Tallman (1998), Orphanides (2000), Croushore and Stark (2000a, 2000b), and Koenig, Dolmas, and Piger (2000).

Table 1
SPF Forecast Error Statistics
 Percent

Error	Full Sample		1996–2000	
	Mean	Std. Dev.	Mean	Std. Dev.
y	.30	1.82	1.93	.84
y'	-.14	1.96	1.68	.98
π	.25	1.36	-.18	.75
u	-.02	.84	-.53	.22
i	-.52	1.53	.10	.79
r	.03	1.29	.30	.50

Note: Full sample period is 1969–2000 except for i and r , which have samples of 1982–2000. **Bold** values indicate that the mean is significantly different from zero at the 5 percent level or better.

otherwise the differences are few.¹⁵ Consequently, the remainder of the analysis focuses on the SPF data because they are available for the longest period.

Several general characteristics of forecast errors are apparent from the figure and table. First, and most obvious, the volatility of errors varies widely. Unemployment rate errors are by far the least variable and GDP errors the most variable. In addition, the volatility of these errors is particularly great in relative terms because the average real interest rate and GDP growth rate are smaller than the averages of inflation, unemployment, and nominal interest rates.

All forecast errors appear to be approximately unbiased, fluctuating more or less evenly around zero over the full sample period, as they should. Most average errors are 0.3 percent or less in absolute value, a relatively small amount for these variables. Only the nominal interest rate average error, at -0.5 percent, is close to being statistically and economically significantly different from zero. But the BC errors suggest that this finding might be an artifact of the shorter sample period.

Although average errors are about zero, some errors are more likely to be one-sided—consecutive positive or negative errors—for extended periods of time. Errors in inflation, unemployment, and to a lesser extent the nominal interest rate often stray significantly above or below zero for many years in a row. Inflation errors were positive in all but one year up to 1981 and then were negative for five straight years.

¹⁵ The significant difference between SPF and BC interest rate errors in 1982 is an anomaly attributable to different data sources and the use of annual averages versus Q4 values (see the Data Appendix for details). Interest rates dropped a lot in the final months of the year so SPF interest rate forecasts (for Q4) made large overprediction errors whereas BC interest rate forecasts made smaller errors because annual interest rates did not decline by nearly as much.

Similarly, unemployment errors were negative from 1983 to 1988 as the unemployment rate was falling. In contrast, errors in GDP (until recently) and real interest rates are more likely to jump up and down across the zero line.

One key reason for this difference is that inflation, unemployment, and nominal interest rates were much more variable. During periods of major structural change, such as the 1970s and early 1980s, forecasters have particular difficulty forecasting the timing of such changes correctly. The left column of Figure 1 shows that inflation and unemployment rate forecasts especially were consistently behind the data while these variables were rising and falling markedly. In contrast, average GDP was relatively stable during this time and forecasters were quite successful at predicting the ups and downs—surprisingly, even better than during the more tranquil period since the mid 1980s.

The figure and table put recent forecast errors into historical perspective. None of the most recent SPF errors are the largest in history. Forecast errors for all variables during the 1970s and early 1980s generally were much larger than those since, primarily because there were more frequent and severe recessions, which are hard to predict.¹⁶ But recent GDP errors are much larger than average. Only the 1983 GDP error was larger than the 1999 error among positive errors, and the 1974 and 1982 errors were the only others larger in absolute value. The GDP errors in the late 1990s are clearly the largest since 1983. Recent errors in unemployment and inflation have been economically significant, but these errors are relatively small in historical perspective.

Table 1 documents that from 1996 to 2000 errors in output growth (1.9 percent) and the unemployment rate (-0.5 percent)—but not inflation, interestingly—were significantly different from zero. This result contrasts sharply with the full sample period, when these errors were roughly zero, and underscores the point that errors may appear to be biased over shorter sample periods. This recent period is also different in that the errors were large during a robust expansion when the macroeconomic data were less variable. In the first half of the sample, large errors occurred primarily during times of economic turbulence, especially severe recessions.

The extended period of large, one-sided GDP errors is troubling. Until the late 1990s, GDP errors had never been one-sided for five consecutive years, much

¹⁶ This point has been made numerous times by McNees (1992, 1994, and 1995).

less *large* and one-sided for that long. Errors in the other variables had been similarly large and one-sided, but not errors in real GDP growth. The one-sided negative errors in the unemployment rate from 1996 to 2000 may

Recent forecast errors are not extraordinarily large in historical perspective, but the extended period of large, one-sided GDP errors is troubling.

be related to the GDP errors during this period, if forecasters rely on Okun's law or some other macroeconomic relationship between output and unemployment. But these 1996–2000 errors for GDP and unemployment are even more puzzling in that they generally have *not* been matched by consistently large, one-sided errors in inflation or interest rates during this period.

Naturally, the question arises: Have forecasting models broken down in the late 1990s and, if so, why? Specifically, are the one-sided GDP and unemployment errors large and long enough to conclude that something has changed fundamentally? Or are these errors statistically similar to historical errors?

Tests for Bias

Figure 2 plots cumulative errors for the SPF average forecasts over two sample periods. The left column shows cumulative errors for the full sample, 1969 to 2000, and the right column cumulative errors for the post-1983 sample. The dashed lines around the cumulative errors indicate the points at which bias is statistically

Figure 1

Average Forecasts and Errors

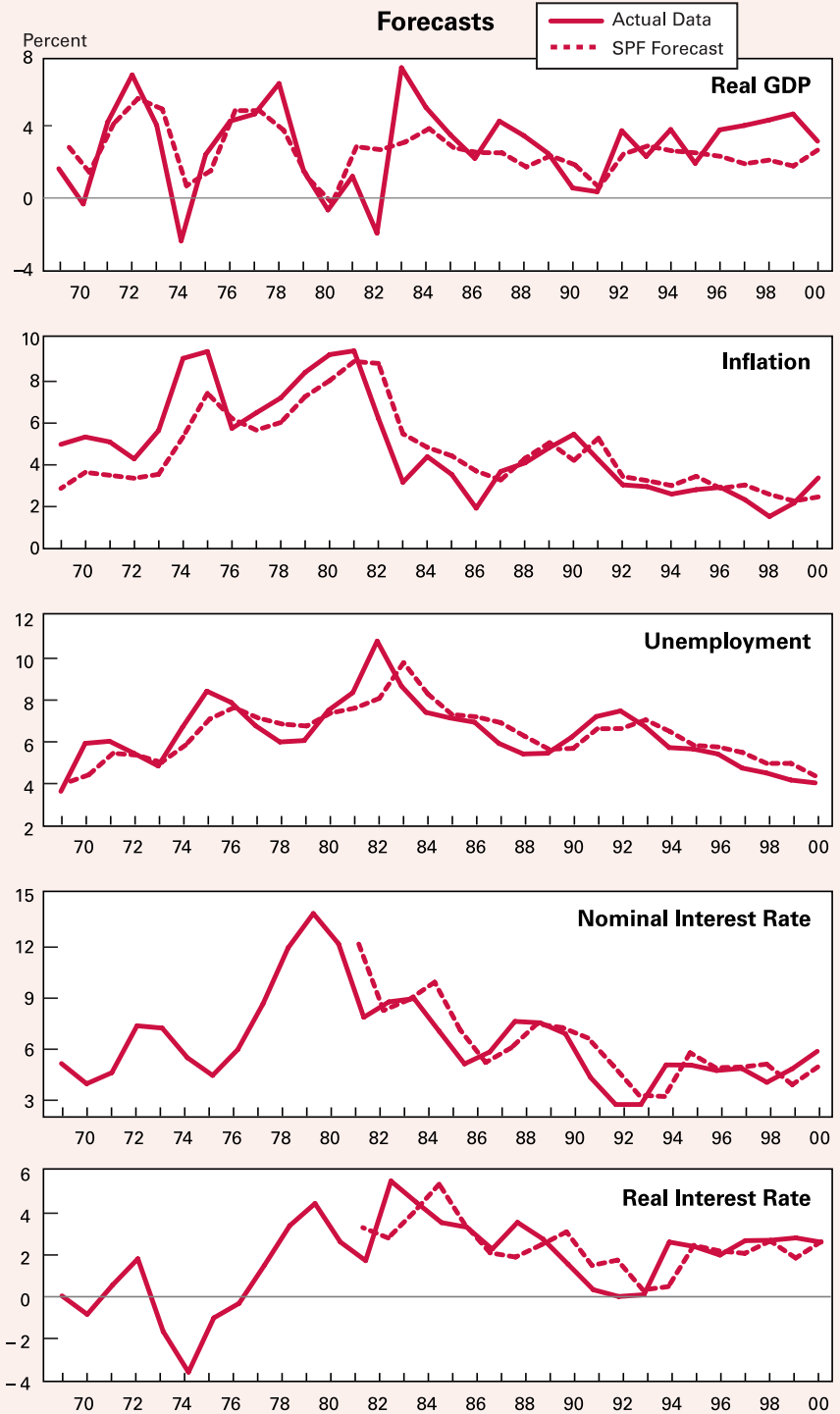
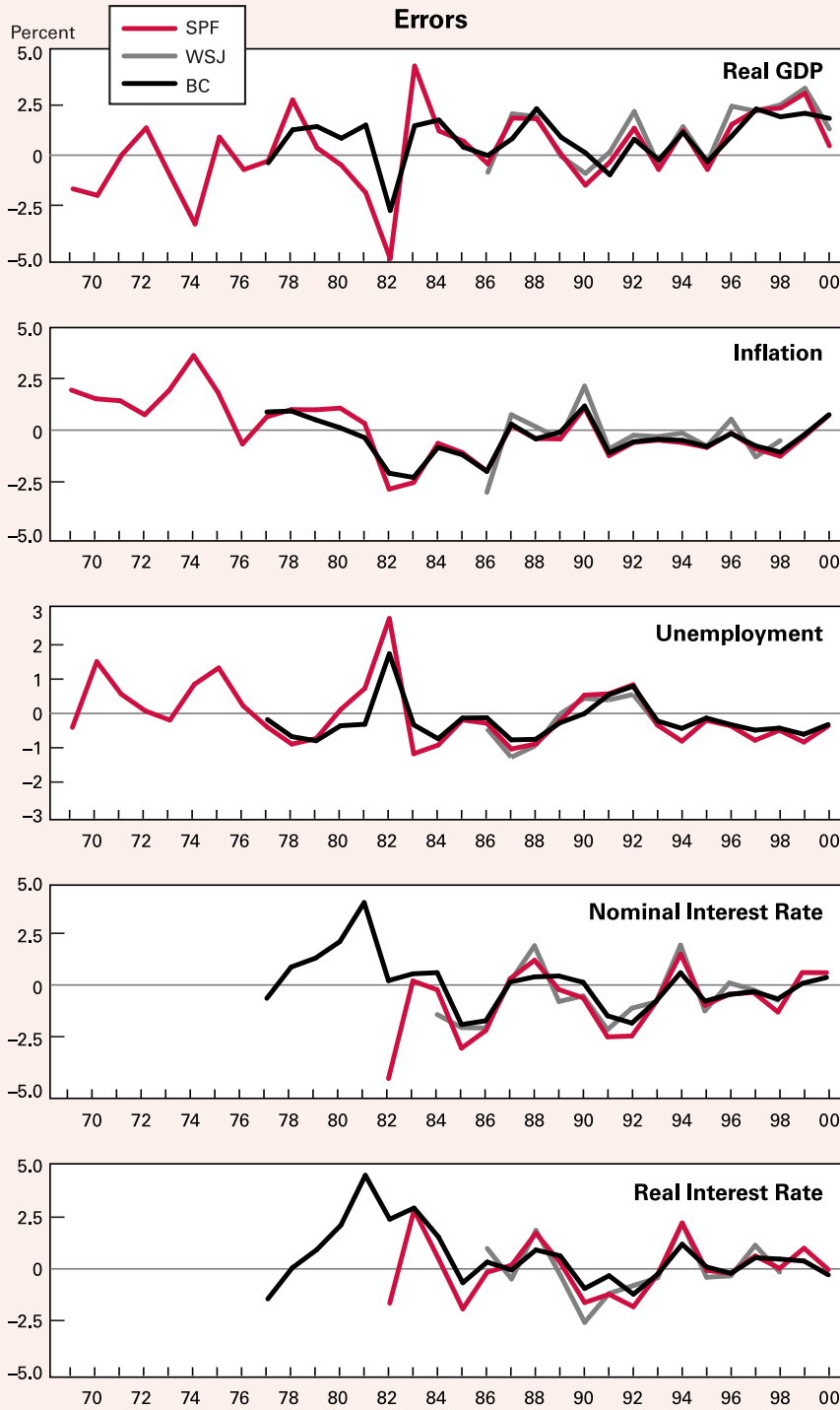


Figure 1, continued

Average Forecasts and Errors



significant at the 5 percent confidence level, which is the basis of the CUSUM test.

The motivation for looking at the post-1983 subsample is that the economy may have experienced significant structural change after the tumultuous period of the 1970s and early 1980s. It is widely believed that monetary policy changed after this period. In addition, McConnell and Perez-Quiros (2000) advanced the hypothesis that there has been a structural shift (reduction) in volatility since 1984 as well. Tests for bias attributable to structural change should not be conducted over periods with multiple breaks, which would be the case for the full sample if breaks occurred in both 1984 and 1996.

Figure 2 provides modest evidence of a statistically significant bias in real GDP forecasts, but only in the post-1983 subsample. The full sample shows no evidence of significant bias in any macroeconomic forecast as of 2000. During the 1970s, inflation forecasts were significantly biased downward (underpredictions) but the forecasts were largely back on track by the 1990s. Over the post-1983 sample, however, real GDP forecasts are biased—but just barely, and it was not until 1999 that one could draw this conclusion confidently. Furthermore, this result is sensitive to the starting date for the cumulative errors. Starting a couple of years earlier or later than 1984, the cumulative errors are not significant. Finally, note that cumulative errors constructed with real-time GDP data show no bias.

In contrast, post-1983 cumulative errors in inflation, unemployment, and nominal interest rates reveal no statistically significant biases. All three cumulative errors are negative (indicating overprediction) and close to their lower bounds, but none are significantly biased. Interestingly—and in contrast to the four other variables—cumulative

errors provide no evidence of bias in real interest rate forecasts, even though these forecasts are merely implied rather than explicit.

Turning to the regression-based evidence, Table 2 reports estimates of bias tests using SPF average forecasts over the full sample and the post-1983 subsample.¹⁷ For each regression (row), the table includes estimates of the intercept and slope. Under the null hypothesis of unbiasedness, these parameters should be zero and one, respectively; significant deviations of either parameter (or both) from these values can lead to a rejection of the null. The p-value from the joint test of this hypothesis appears in the final column and indicates the level of confidence at which the hypothesis can be rejected.

The regressions yield qualitatively similar results to the CUSUM tests—most forecasts of macroeconomic variables are unbiased. Over the full sample, the p-values generally do not come close to conventional levels of significance for rejection except for the nominal interest rate, which clearly is biased, but these data are not available before 1982. However, like the CUSUM tests, the regressions show some evidence of bias in the post-1983 subsample. GDP forecasts are biased relative to current data (but not real-time data) over this subsample, and unemployment forecasts are on the borderline. But this result also is sensitive to the starting period of the sample, as it is with the CUSUM test.

Overall, the statistical evidence of bias is rather weak and not very

¹⁷ Estimates with the BC average forecasts over the 1977–2000 period are very similar. The only two substantive differences are that the BC forecasts of GDP are biased relative to the current data and the BC coefficients in the nominal interest rate regression are quite different (intercepts with different signs). The former is not attributable to the shorter sample period, as SPF forecasts of GDP for the 1977–2000 period are unbiased.

Figure 2

Cumulative Errors for SPF Average Forecasts

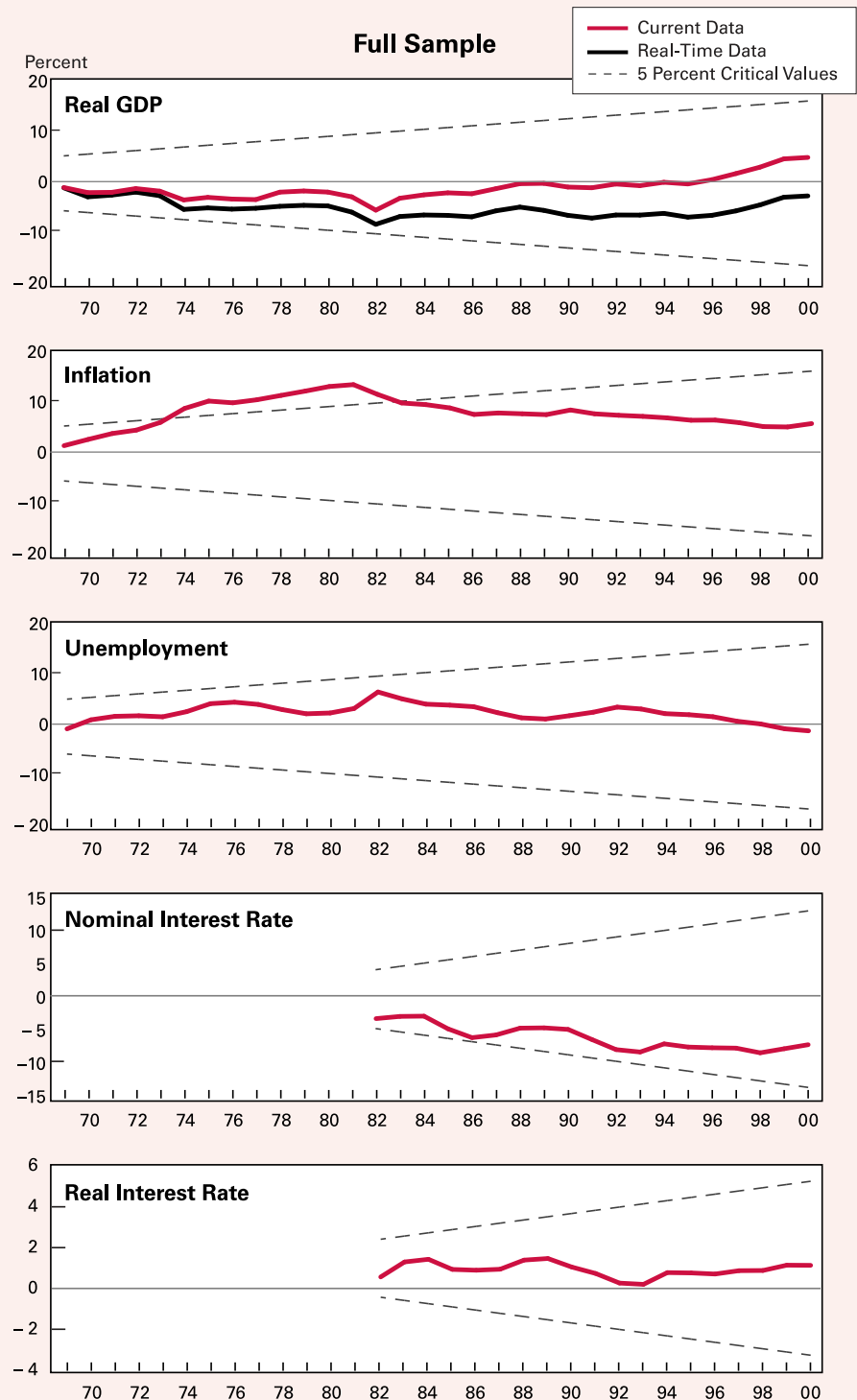
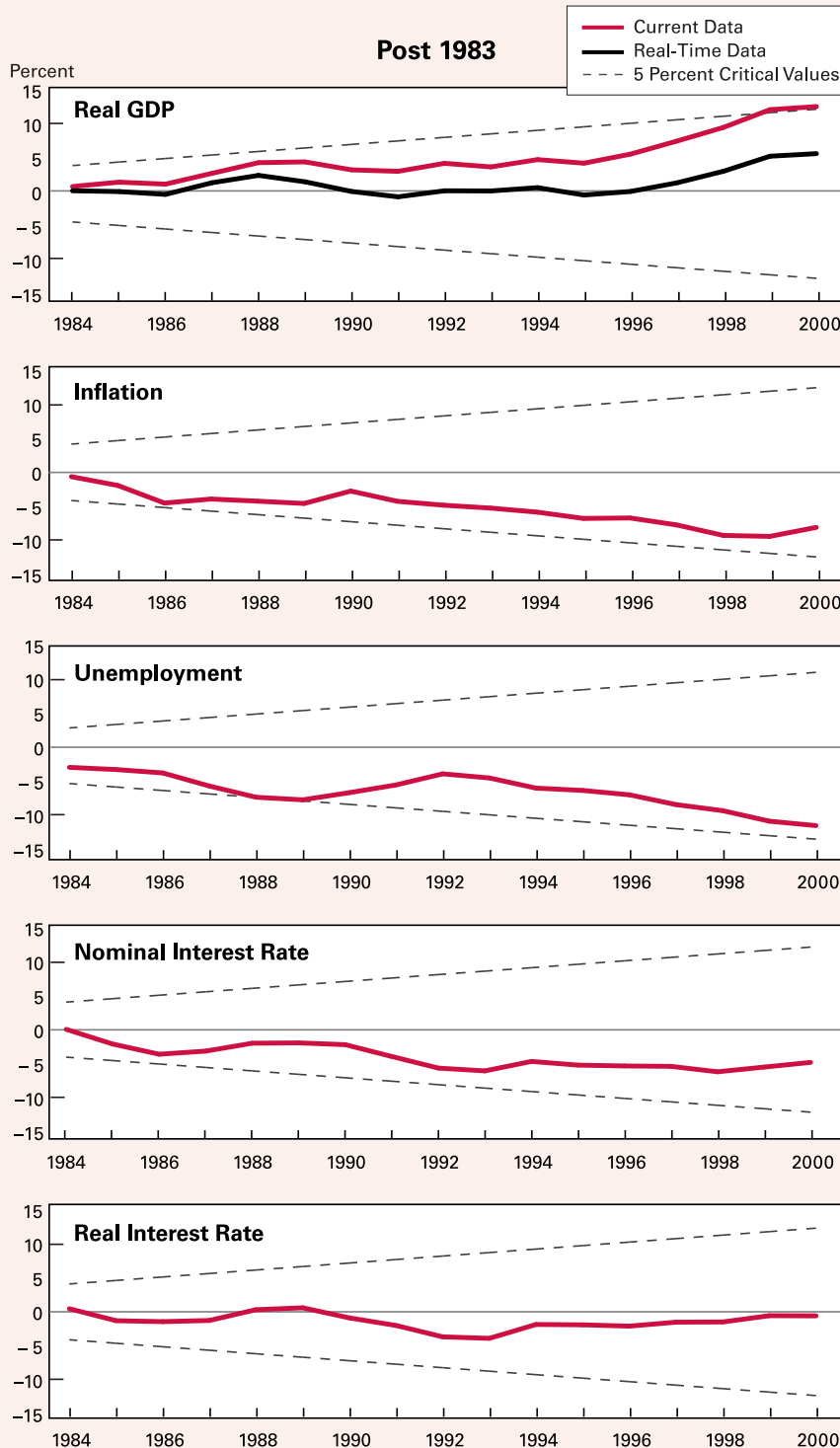


Figure 2, continued

Cumulative Errors for SPF Average Forecasts



robust. This result is somewhat surprising in light of the fact that many previous tests have found substantial biases in average forecasts, particularly inflation forecasts. Apparently,

Most forecasts of macroeconomic variables are unbiased over longer periods of time.

moderately large errors over medium-term periods tend to average out over longer periods of time.

Tests for Efficiency

Regression-based tests of efficiency for average forecasts require a specification of Z_t , the explanatory variables that might improve forecasts. Individual forecasters making forecasts in period t of output, inflation, unemployment, and interest rates surely know their own forecasts plus the lagged data for each variable. If so, then $Z_{it} =$

$[\hat{y}_{it}, \hat{\pi}_{it}, \hat{u}_{it}, \hat{r}_{it}, \hat{y}_{i,t-1}, \hat{\pi}_{i,t-1}, \hat{u}_{i,t-1}, \hat{r}_{i,t-1}]$ is a reasonable choice, and none of these variables should significantly improve the fit of the regression. Although there is no such thing as the “average forecaster,” I make the analogous assumption for average forecasts.¹⁸ With only 32 annual observations of average forecast data, and fewer for individual forecasters, it is not possible to include all Z variables simultaneously so they are added to efficiency regressions one at a time. (Only one significant variable is required to reject efficien-

¹⁸ In fact, most individual forecasters probably know the average forecast of other forecasters as well. But this hypothesis is more speculative and requires more investigation.

Table 2
Regression Tests of Bias for SPF Aggregate Forecasts

Variable	Sample	β_0	β_1	\bar{R}^2	p-value
\tilde{y}	1969–2000	-.07 (.77)	1.13* (.25)	.39	.59
	1984–2000	.93 (1.18)	.97* (.45)	.19	.04
\tilde{y}^r	1969–2000	-.83 (.83)	1.24* (.26)	.41	.61
	1984–2000	.32 (1.40)	1.06* (.53)	.16	.42
$\tilde{\pi}$	1969–2000	.99 (.95)	.86* (.18)	.76	.57
	1984–2000	.14 (.76)	.87* (.20)	.48	.19
\tilde{u}	1969–2000	-.02 (.78)	1.00* (.12)	.68	.99
	1984–2000	-.24 (.86)	.99* (.14)	.76	.10
\tilde{i}	1982–2000	2.28* (.91)	.57* (.13)	.49	.01
	1984–2000	1.81 (1.07)	.64* (.17)	.45	.08
\tilde{r}	1982–2000	.98 (.69)	.62* (.25)	.22	.33
	1984–2000	.90 (.58)	.62* (.26)	.32	.23

Note: Standard errors are in parentheses. * indicates the coefficient is significant at the 10 percent level or better. The p-value is from the F-test; **bold** indicates rejection of unbiasedness at the 10 percent level or better. The $\tilde{\pi}$ regressions include an AR(1) correction for serial correlation estimated by maximum likelihood with a grid search for the global maximum.

cy.) Not all forecasters report data for all five macroeconomic forecasts.

Table 3 reports results from efficiency tests of SPF average forecasts over the full sample. The table format is somewhat nonstandard in that columns do not represent results from a single regression; rather, each number represents an estimate of the β_2 parameter on Z_t from a separate regression (with standard errors in parentheses).¹⁹ The bottom portion of the table provides information about the econometric importance of Z_t . Bias \bar{R}^2 is from Table 2, Max. $\Delta\bar{R}^2$ is the maximum increase in fit from incorporating Z_t , and Max. Z_t is the variable that yielded the maximum fit. Bold estimates indicate a significant rejection of efficiency.

All forecasts fail tests for efficiency except the real interest rate forecast. For each of the four other fore-

¹⁹ The real interest rate identity makes some of the inflation and interest rate estimates identical.

casts, at least three measures of Z_t are significant and efficiency is rejected. Moreover, the magnitude of the estimates is often quite large. Not surprisingly, these additional variables make economically significant contributions to the fit of the data. For GDP and unemployment, the fit increases by about 25 percent; for inflation and interest rates, the fit improves by more than 5 percent.

Two interesting patterns emerge from an examination of the variables that improve each forecast. First, data on inflation and nominal interest rates (both forecasts and lagged data) improve forecasts of GDP and unemployment, and the improvements are rela-

Table 3
Regression Tests of Efficiency for SPF Aggregate Forecasts

Z	Dependent Variable				
	y_t	π_t	u_t	i_t	r_t
\tilde{y}_t		-.23 (.15)	-.10 (.11)	.51 (.44)	.64 (.44)
$\tilde{\pi}_t$	-.38* (.17)		.33* (.09)	-.14 (.40)	-.11 (.21)
\tilde{u}_t	.12 (.27)	-.65* (.24)		.12 (.28)	.24 (.23)
\tilde{i}_t	-.45* (.19)	.04 (.19)	.28* (.10)		-.11 (.21)
\tilde{r}_t	-.27 (.43)	.04 (.19)	.10 (.19)	-.14 (.40)	
y_{t-1}	.01 (.15)	.02 (.08)	-.18* (.05)	.12 (.14)	.05 (.15)
π_{t-1}	-.35* (.13)	-.42* (.21)	.29* (.06)	-.34 (.25)	-.16 (.17)
u_{t-1}	.09 (.25)	-.48* (.17)	-.25 (.32)	.07 (.22)	.19 (.19)
i_{t-1}	-.33* (.13)	.38* (.19)	.19* (.06)	.79 (.89)	-.10 (.20)
r_{t-1}	-.01 (.18)	.42* (.12)	-.05 (.08)	.38 (.24)	.42 (.37)
Bias \bar{R}^2	.39	.76	.68	.49	.22
Max. $\Delta\bar{R}^2$.10	.05	.15	.03	.05
Max. Z_t	π_{t-1}	u_{t-1}, r_{t-1}	π_{t-1}	r_{t-1}	\tilde{y}_t

Note: Each entry reflects an estimate of β_2 from a separate regression, with standard errors in parentheses. * indicates that the coefficient estimate is significant at the 10 percent level or better, and **bold** estimates indicate that efficiency (joint test of $[\beta_0, \beta_1, \beta_2] = [0, 0, 0]$) is rejected at the 10 percent level or better. Missing entries indicate the estimates are not applicable. Bias \bar{R}^2 is from the bias regressions in Table 1, Max. $\Delta\bar{R}^2$ indicates the maximum increase for any Z in the column, and Max. Z indicates which variable achieved the maximum. The sample periods are 1969 to 2000 except for regressions involving \tilde{i}_t and \tilde{r}_t , which are 1982 to 2000. The $\tilde{\pi}$ regressions include an AR(1) correction for serial correlation estimated by maximum likelihood with a grid search for the global maximum.

tively large. Second, data on unemployment and interest rates improve forecasts of inflation, but by a more modest amount.

Data on inflation and nominal interest rates (both forecasts and lagged data) improve forecasts of GDP and unemployment, and the improvements are relatively large.

Apparently not all facets of inflation and interest rates are being incorporated into average GDP forecasts. The first column indicates that average GDP forecasts might be more accurate if they were adjusted to account for a strong inverse correlation between GDP forecast errors and inflation or interest rates. In other words, when inflation and interest rates are unusually high, GDP forecasts should be lower, and vice versa. The estimates for inflation and interest rates, forecasts and lagged data, are all large and similar, though inflation has a stronger impact.²⁰ A decrease in inflation or interest rates of 1 percentage point, either in the lagged data or in the current forecast, should raise the current GDP forecast by about 0.3 to 0.5 percentage point. This analysis is essentially the same for unemployment except that the estimates are positive and smaller in absolute value.

Inflation forecasts also fail to completely incorporate all information, especially on unemployment and interest rates. The second column indicates that average inflation forecasts might fit the data better if they were adjusted to account for the strong inverse correlation between inflation errors and unemployment or the strong positive correlation between inflation errors and interest rates. Apparently, when unemployment is unusually low or interest rates are unusually high, average inflation forecasts should be higher. These estimates are even larger than for GDP and unemployment, except for the insignificant estimates on interest rate forecasts.²¹

²⁰ Unreported GDP regressions that include both inflation and interest rates together show little additional explanatory power. Apparently, the information helpful to GDP forecasts is common to both explanatory variables.

²¹ For inflation, unreported regressions with multiple variables reveal independent contributions from unemployment and interest rates.

An explanation for this correlation may be associated with monetary policy. Perhaps private forecasters have inaccurate assessments and forecasts of the Federal Reserve's interest rate and inflation targets, or inaccurate estimates of its response to macroeconomic shocks. Alternatively, perhaps the Fed is better at predicting inflation or uses a different model to forecast inflation. These interesting hypotheses merit further investigation. In any case, the explanation for inefficiency can only be found in the details of the macroeconomic forecasting models.

Evidence on efficiency in interest rate forecasts is mixed. Efficiency is rejected for every variable in the nominal interest rate regressions, but none of the estimates are significant. This means that the rejection is primarily attributable to problems in the other two parameter estimates rather than to meaningful contributions by the alternative variables. In contrast, efficiency is not rejected for any variables in the real interest rate regressions. The shorter sample for interest rates seems to limit the conclusions that can be drawn confidently.

Tests for Correlation

Table 4 reports results of regression tests of correlation between current and lagged forecast errors. Such correlation should not exist because, if it did, forecasters could use it to improve their forecasts by revising their models or by obtaining better econometric estimates of the model parameters.

The first panel reports regressions of each error against its own lag only, called an AR(1). The next two panels report multivariate regressions of all lagged errors, in search of cross-correlation. The panel with GDP, inflation, and unemployment excludes interest rates and thus covers the full sample period. The panel with inflation and the nominal interest rate cannot include the real interest rate as well, so the final panel reports the real rate estimate from a separate restricted regression.

The AR(1) results indicate that most average forecast errors are not correlated with their own lags (serially correlated). The exception is inflation, which is highly serially correlated; the unemployment error shows some serial correlation but is not quite significant. The lagged inflation error is also highly correlated with the GDP and unemployment errors over the full sample, but none of the other cross-correlations are significant. Moreover, this correlation appears to be associated primarily with the high inflation period of the 1970s and early 1980s. Over the shorter sample

Table 4

Regression Tests of Correlation for SPF Aggregate Forecast Errors

Sample	Independent Variables	Forecast Error				
		ϵ_t^Y	ϵ_t^P	ϵ_t^U	ϵ_t^I	ϵ_t^E
1969– 2000	AR(1) only	.09 (.18)	.57** (.15)	.27 (.18)	.10 (.21)	.01 (.24)
	ϵ_{t-1}^Y	.05 (.26)	-.06 (.16)	-.15 (.13)		
	ϵ_{t-1}^P	-.60** (.24)	.59** (.15)	.31** (.09)		
	ϵ_{t-1}^U	.21 (.55)	-.47 (.35)	-.03 (.22)		
	\bar{R}^2	.12	.34	.36		
1982– 2000	ϵ_{t-1}^Y	-.12 (.31)	-.04 (.19)	.01 (.12)	-.05 (.32)	-.09 (.31)
	ϵ_{t-1}^P	-.70 (.44)	.05 (.27)	.32* (.17)	-.46 (.45)	-.51 (.43)
	ϵ_{t-1}^U	-.07 (.85)	-.19 (.52)	.17 (.34)	.04 (.89)	.23 (.85)
	ϵ_{t-1}^I	-.08 (.41)	.12 (.25)	.02 (.16)	.31 (.42)	.19 (.40)
	\bar{R}^2	.08	.02	.05	-.19	-.12
	ϵ_{t-1}^E	.26 (.39)	.04 (.21)	-.13 (.16)	.37 (.36)	.33 (.35)
	\bar{R}^2	-.07	.06	-.15	-.12	-.09

Note: ** indicates significance at the 5 percent level and * indicates significance at the 10 percent level.

period, 1982 to 2000, forecast errors are much less correlated with the lagged inflation error—especially the inflation error itself—and only the correlation with the unemployment error is significant.

Correlation between forecast errors and lagged inflation errors may provide information that could be used to improve forecasts. Certainly the inflation forecasts could be improved by adding greater persistence. Exploiting the cross-correlation with GDP and unemployment is more difficult, however. Apparently, the average GDP (unemployment) forecast overlooks the fact that when inflation is overpredicted last period, current GDP growth (unemployment) will be higher (lower) than forecast this period. The explanation for this correlation also may be associated with monetary policy.

Implications and Applications

The statistical analysis in this section points toward the overall conclusion that average macroeco-

nomical forecasts generally are unbiased but inefficient, with correlations among forecast errors that apparently are not exploited. One possible interpretation of this conclusion is that recent forecasting errors do not reflect a breakdown in forecasting models, but rather a confluence of macroeconomic conditions that magnified preexisting inefficiencies and caused forecasts to run off track temporarily. This latest episode of errors also may have increased the probability of detecting the forecasting problems.

This interpretation seems broadly consistent with recent GDP and unemployment errors, but perhaps not the relatively smaller inflation errors. During the late 1990s, inflation and nominal interest rates were low by historic standards and forecasters consistently overpredicted both variables. The efficiency and correlation tests suggest that taking account of the cor-

relation between GDP (and unemployment) forecast errors and these variables might help explain the large GDP forecast errors.

However, recent macroeconomic conditions do not seem to help explain inflation forecast errors. On one hand, the efficiency tests suggest that unusually low nominal interest rates should produce *lower* inflation than was forecast. On the other hand, the tests suggest that unusually low unemployment should produce *higher* inflation than was forecast. This tension probably is related to the errors made in Phillips curve relationships over this period.

To quantify the extent to which recent macroeconomic conditions interacted with forecast inefficiencies and error correlation, I constructed various adjusted average forecasts that try to correct for inefficiencies and correlation. The adjusted forecasts were obtained from the fitted values of efficiency regressions that include selected measures of Z_t and/or lagged errors that were found to be significant. To ensure that the adjustments are not unduly

influenced by the latest developments, the efficiency regressions are estimated from 1969 to 1995 (the results are still significant). Then the 1969–95 estimates are used to construct out-of-sample (1996 to 2000) adjusted forecasts. Adjusted forecasts using the efficiency regressions estimated over the full sample period (1969–2000) are also reported for comparison. If inefficiency and correlation were unchanged over the full sample period, including the last five years in the regressions would give a more accurate adjustment to the forecasts.

Table 5 summarizes the impact of the efficiency adjustments. Adjusting for forecast inefficiencies significantly improves recent GDP forecasts. Adding only the lagged nominal interest rate to the GDP regression reduces the average 1996–2000 GDP error of 1.9 percent by one-fourth (1969–1995 adjustment) to one-half (1969–2000 adjustment). Moreover, the variance of the GDP forecast error is reduced by about one-fourth over the full sample. Adding only the lagged inflation error yields more modest reductions in average errors and error variances. Together, these two variables reduce the average 1996–2000 error by up to two-thirds and the full-sample variance by one-third. These are economically significant improvements.

In contrast, adjusting for forecast inefficiencies yields mixed results for inflation forecasts. Adding the lagged real interest rate essentially eliminates the relatively small average 1996–2000 error of –0.2 percent, and it cuts the full-sample variance of the inflation error in half. However, adding lagged unemployment or the lagged inflation error makes the forecast error worse by inducing more bias during the 1996 to 2000 period, even though these variables also reduce the full-sample error variance. This result suggests that there may indeed be a fundamental break in the structural determination of inflation during recent years.

To summarize, it appears that average macroeconomic forecasts could be improved to reduce the puzzling errors encountered during the late 1990s. Rather than adding information in an ad hoc manner, a better way to make those adjustments is to restructure the econometric models underlying these average forecasts. That requires working knowledge of the models.

Table 5
SPF Aggregate Forecast Errors Adjusted for Inefficiency
Percent

Error	Adjustment Variables	1996-2000 Average Errors			Reduction in Error Variance	
		Actual	Adjusted by Regression		1969–1995	1969–2000
			1969–1995	1969–2000		
$\tilde{\epsilon}^y$	i_{t-1}	1.9	1.4	1.0	23	25
$\tilde{\epsilon}^y$	$\tilde{\epsilon}_{t-1}^\pi$	1.9	1.7	1.2	6	21
$\tilde{\epsilon}^y$	$i_{t-1}, \tilde{\epsilon}_{t-1}^\pi$	1.9	1.2	.7	23	35
$\tilde{\epsilon}^\pi$	r_{t-1}	–.18	–.00	–.01	50	51
$\tilde{\epsilon}^\pi$	u_{t-1}	–.18	–.81	–.45	45	49
$\tilde{\epsilon}^\pi$	r_{t-1}, u_{t-1}	–.18	–1.20	–.68	43	47
$\tilde{\epsilon}^\pi$	$\tilde{\epsilon}_{t-1}^\pi$	–.18	–.37	–.25	41	41

However, it is possible that some individual forecasters may have had better success during the late 1990s. If so, we might be able learn something about the unusual average forecast errors from the forecasts and errors of successful individual forecasters.

IV. Individual Forecasts

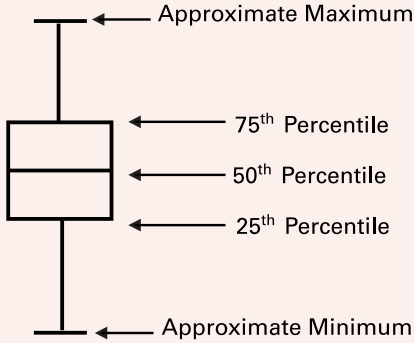
This section briefly examines the forecast performance of individual forecasters from the SPF and WSJ. A thorough examination of individual forecast properties is left for future research. The analysis is largely parallel to that of the previous section. It presents historical data on forecasts and forecast errors and reports tests for bias, efficiency, and correlation for individual forecasters. Finally, it provides a brief summary assessment of individual forecasters' performance.

Figure 3 plots actual data against the forecasts and forecast errors of individual SPF forecasters for the five macroeconomic variables.²² This figure is analogous to Figure 1 except that it reflects an annual summary of all forecasters through a device called a box plot. The box plot indicates the major percentiles of the distribution across forecasters that year—see Illustration 1 for details. Note that specific forecasters are not identified in the figure, and any particular forecaster could be at the top of the box plot in one year and at the bottom the next year.

²² The WSJ forecasters are not included because their sample period differs from that of the SPF forecasters and the WSJ forecasts are somewhat more diverse, so when the WSJ forecasters are included the distribution appears to change significantly. Qualitatively, the version of Figure 3 with WSJ forecasters is the same but the height of the box plots (dispersion of forecasts) is somewhat greater on average.

Illustration 1

Box Plot Definitions



Note: A small number of points may be outside the Approximate Maximum and Minimum.

The broad patterns in the forecast and forecast error distributions are essentially the same as those for the average SPF forecasts. GDP errors tend to be the most volatile, and the volatility of the data, forecasts, and forecast errors is considerably lower in the post-1984 period. Recent errors are relatively small in historical perspective but tend to be one-sided (especially for GDP). Most of the forecast error box plots encompass zero, which means that at least some forecasters were “right” in most—but clearly not all—years. Forecasters underpredicted inflation throughout the 1970s, but real interest rate forecasts were surprisingly and consistently accurate.

The most striking feature of Figure 3 is that essentially *all* individual forecasters underpredicted GDP throughout the late 1990s and, until recently, the errors got worse. For the four years from 1996 to 1999, the GDP error box plots lie above zero and the median increased each year. Only in 2000 did the median error decline, and a minority of SPF forecasters made errors close to zero (none of the WSJ forecaster errors were zero in

Figure 3

Individual SPF Forecasts and Errors

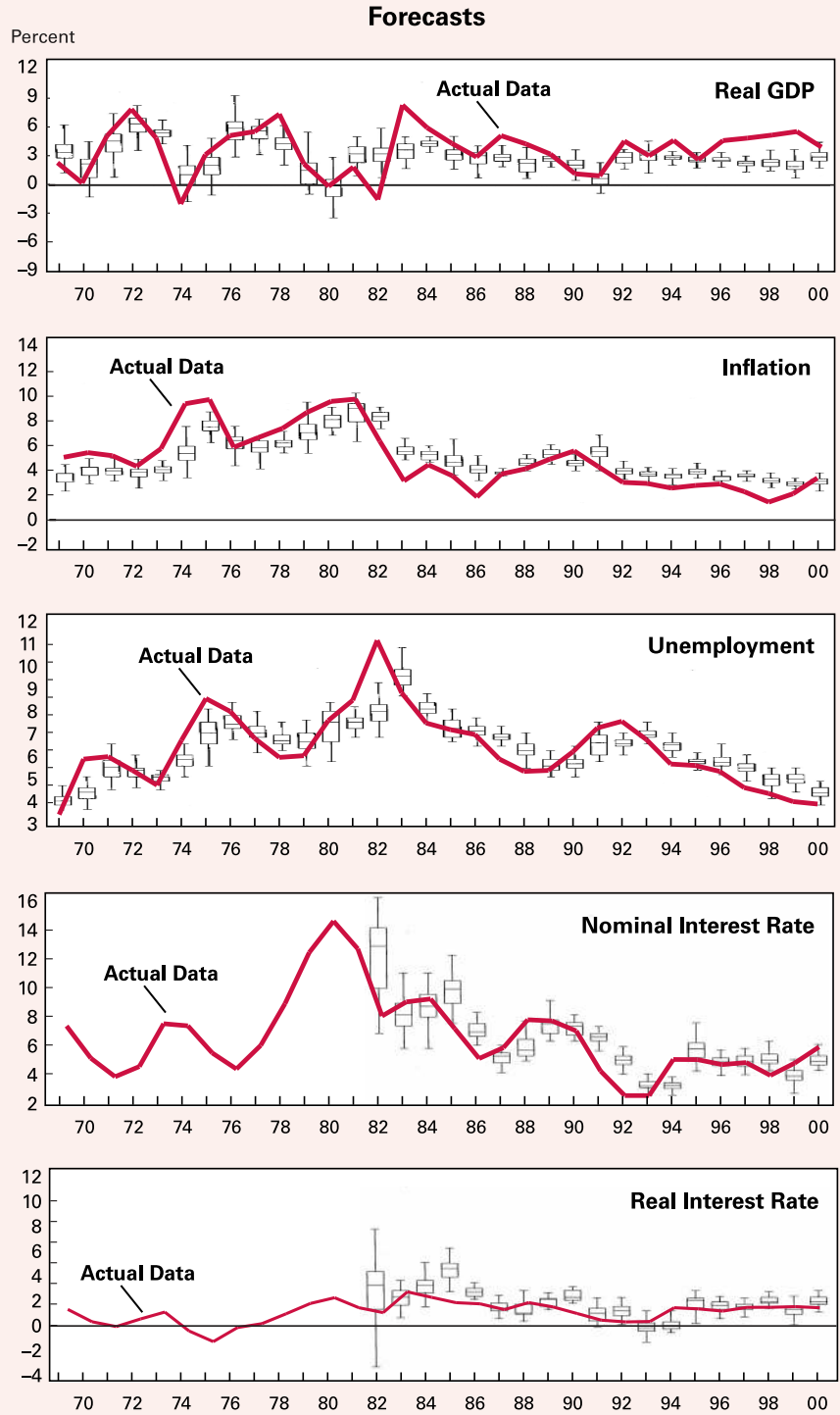
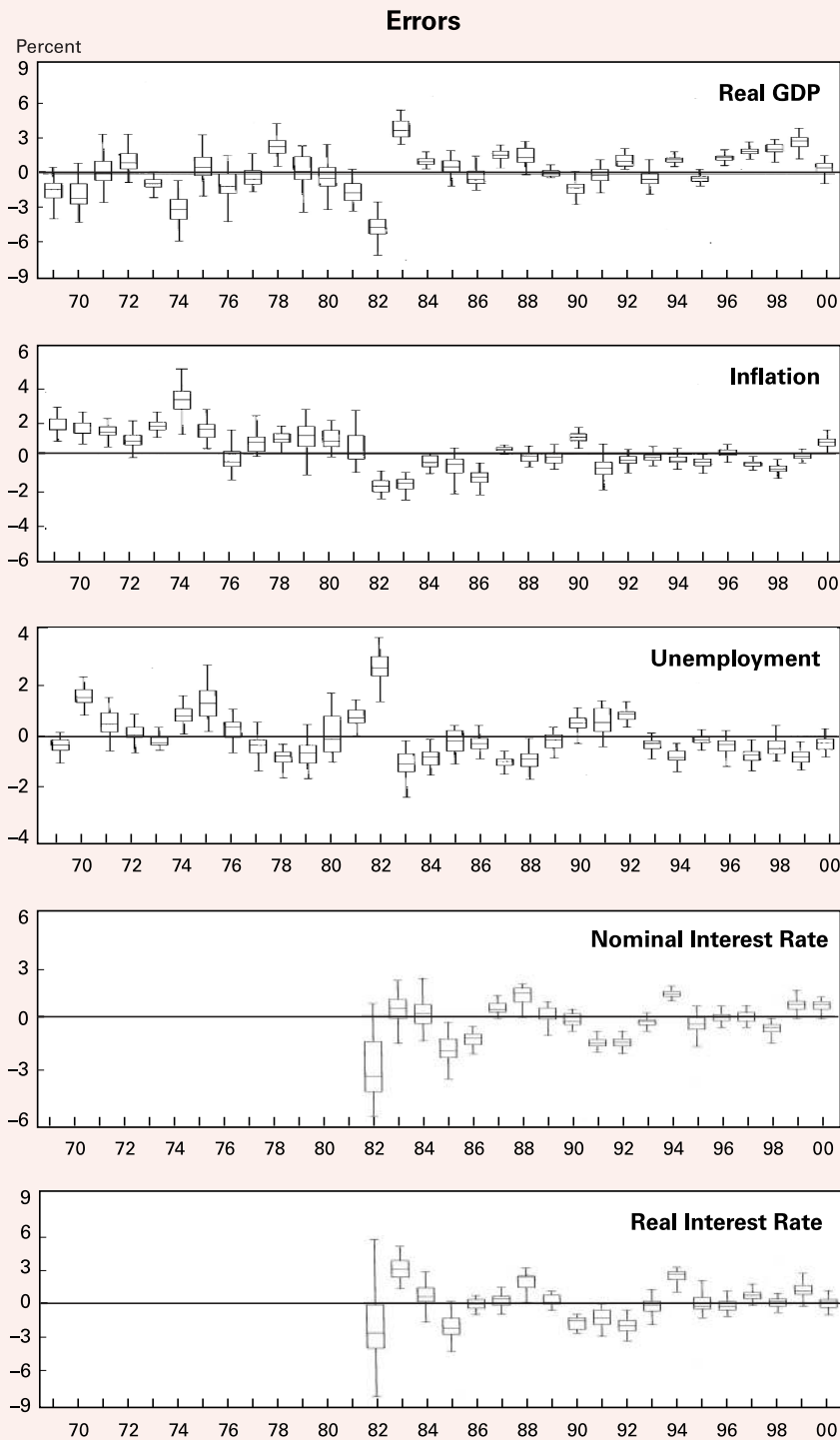


Figure 3, continued

Individual SPF Forecasts and Errors



2000). This improvement in 2000 was the joint result of forecasters increasing their GDP forecasts and GDP data declining unexpectedly at the end of the year. Prior to 1996, the box plots failed to encompass zero in only six years, and in those episodes the median was never the same sign for more than two consecutive years.

Individual unemployment forecasts in the late 1990s also tended to be one-sided but the distributions were not as bad as for GDP. Most unemployment forecast errors, including the median, were negative from 1996 to 2000. However, a small minority of forecasts was relatively accurate each year, and the distribution of forecast errors was relatively small in historical perspective.

In general, inflation forecast errors were not particularly large during this period except in 1997 and 1998. During these two years essentially all forecasters overpredicted inflation. This relatively short string of errors is less troubling, as the inflation errors disappeared in 1999 and turned positive in 2000, largely because of unexpected energy price increases. Furthermore, these inflation errors are smaller and less persistent than those prior to the mid 1980s.

Overall, Figure 3 indicates that there is little hope of gleaning an explanation for the recent large and one-sided forecast errors from individual forecasters. Because no one consistently forecasted GDP and unemployment accurately over the period 1996 to 2000, we cannot obtain clues to the forecast error biases from forecasters who "got it right." Nevertheless, it remains of interest to test individual forecasts, in order to gain a better understanding of the average forecast error properties and to ascertain whether some forecasters offer relatively better forecasts.

The remainder of this section reports the results of these tests for individual SPF and WSJ forecasters.

Table 6

Regression Tests of Bias for SPF and WSJ Individual Forecasts

	Forecast				
	\bar{y}	$\bar{\pi}$	\bar{u}	\bar{i}	\bar{r}
Unbiased					
Total Forecasts	17	22	15	4	11
Early Group	12	1	14	1	2
Later Group	5	21	1	3	9
β_0 mean	.51 (.35)	.51 (1.31)	.65 (.76)	1.97 (.37)	.92 (.40)
β_1 mean	.78 (.37)	.85 (.40)	.91 (.12)	.65 (.09)	.58 (.25)
p-value mean	.41 (.22)	.53 (.30)	.66 (.27)	.31 (.31)	.40 (.13)
Biased					
Total Forecasts	20	16	2	11	1
Early Group	3	15	1	3	0
Later Group	17	1	1	8	1
β_0 mean	2.34 (1.19)	2.52 (1.03)	-.20 (4.08)	2.53 (1.64)	1.26 n.a.
β_1 mean	.45 (.49)	.65 (.16)	.99 (.64)	.51 (.24)	.46 n.a.
p-value mean	.03 (.02)	.02 (.02)	.07 (.04)	.03 (.03)	.09 n.a.

Note: Numbers in parentheses are standard deviations of the parameter estimates across all forecasters. Sample periods vary by forecaster with a minimum of 10 and maximum of 21 observations. Early Group contains forecasters whose forecasts begin before 1984 (mostly WSJ) and the Later Group contains forecasters whose forecasts begin after 1983 (mostly SPF).

Only forecasters with 10 or more annual observations are included in the tests.²³ Most of these forecasters have 11 to 14 observations, and the maximum number of observations is 21. Most SPF forecaster data begin before 1984 and most WSJ forecaster data begin after 1983, an important sample difference that is noted in the results. Not all forecasters report all variables every year, so the coverage varies across macroeconomic variables. Data on GDP and inflation forecasts are the most widely available.

Forecast Tests

Table 6 summarizes the results of the individual forecaster bias regressions. The table reports the number of individual forecasters with unbiased and biased forecasts for each variable, as well as the numbers for two groups distinguished by whether their forecasts begin before 1984 (Early Group) or after 1983 (Later

Group). The Early Group includes mostly SPF forecasters and the Later Group mostly WSJ forecasters, but some of each forecaster type are in each group. The table also lists the average parameter estimates and the standard deviation (not standard error) of the estimates in parentheses.

The numbers of biased and unbiased individual forecasts are roughly the same for GDP and inflation forecasts, but the test results are closely linked to the sample period. Most GDP forecasts in the Later Group are biased (17 out of 22), whereas most GDP forecasts in the Early Group are unbiased (12 out of 15). The difference is even more pronounced for inflation. Virtually all inflation forecasts in the Later Group (21 out of 22) are unbiased, whereas virtually all inflation forecasts in the Early Group (15 out of 16) are biased.²⁴

Most unemployment forecasts are unbiased, but this result may be sample-dependent as well. Data availability limits the unemployment rate forecasts to primarily SPF forecasters whose data begin prior to 1984. Most nominal interest rate forecasts are biased and most real interest rate forecasts are unbiased, but most interest rate forecasts begin after 1983. For all variables, the average parameter estimates are comparable to those of the average forecasts.

Table 7 summarizes the results of the individual forecaster efficiency regressions. Each number in the table represents the percentage of forecasters whose forecast (row) fails efficiency when each element of Z_{it} is added to the regression. For example, the first number in the first row indicates that efficiency can be rejected for 46 percent of all individual GDP forecasts because their own inflation forecast contains useful information for predicting GDP. Thus, high numbers indicate greater rejections of efficiency in individual forecasts. The table reports these percentages for all forecasters and for the two groups of forecasters divided by sample period.

Widespread rejection of efficiency is seen among forecasters for all variables except the real interest rate.

²³ It is possible that choosing only forecasters with long sample periods imparts a sample selection bias. For example, forecasters with less than 10 observations might be biased and/or inefficient and thus go out of business before they survive 10 years. Unfortunately, however, the forecast surveys do not track the reasons for entry and exit of forecasters, nor do they provide a valid statistical representation of all forecasters, so it is not possible to quantify potential sample selection bias very well.

²⁴ The individual bias results are consistent with the average bias results reported in Table 2 for forecasts beginning after 1983. However, apparently the biases in forecasts over short subsamples average out over the full sample. Also, note that the average forecast results can differ from the individual results because the average forecasts include forecasts from individuals with less than 10 observations.

For all forecasts, at least one alternative variable induces a rejection of efficiency for at least half of all individual forecasters. Perhaps the most remarkable result is that efficiency is rejected for four out of five individual inflation and interest rate forecasts because forecasters apparently did not properly incorporate *their own* unemployment forecasts. Information in unemployment forecasts leads to a rejection of efficiency for one-half of the real interest rate forecasts as well. This result indicates that most individual forecasters could make some improvements in their inflation and interest rate forecasts with very little difficulty or cost by simply adjusting for the correlation with their unemployment forecasts. They also might be able to make significant improvements by redesigning their forecasting models to eliminate this inefficiency.

The efficiency results vary across forecaster groups in two ways. First, the vast majority of rejections of efficiency for GDP forecasts occur in forecasts from the Later Group. All variables lead to the rejection of efficiency in GDP forecasts for at least three-fifths of the forecasts from the Later Group, but fewer than one in four alternative variables leads to a rejection of efficiency in GDP forecasts from the Early Group. Second, the vast majority of efficiency rejections in inflation forecasts occur in forecasts from the Early Group, while inflation forecasts from the Later Group are rarely inefficient.

Finally, Table 8 reports the results of tests for correlation among individual forecast errors by regressing contemporaneous errors on lagged errors. Analogous to Table 7, entries in this table are percentages of forecasters whose coefficient estimates are significant. The diagonal entries reflect the AR(1) estimates. The off-diagonal entries are estimates from regressions of the contemporaneous errors on each of the lagged errors, one lagged error at a time.

The majority of individual errors do not exhibit correlation, but many individual errors do. Only one-third of the inflation errors exhibit autocorrelation (the

Table 7

Regression Tests of Efficiency for SPF and WSJ Individual Forecasts
Percent Rejecting Efficiency

Variable	Z_{it}									
	\tilde{y}_{it}	$\tilde{\pi}_{it}$	\tilde{u}_{it}	\tilde{r}_{it}	\tilde{r}_{it}	$y_{i,t-1}$	$\pi_{i,t-1}$	$u_{i,t-1}$	$i_{i,t-1}$	$r_{i,t-1}$
<i>All Forecasters</i>										
\tilde{y}_t		46	31	69	50	43	59	51	62	43
$\tilde{\pi}_t$	38		82	23	17	30	27	35	35	38
\tilde{u}_t	6	29		40	0	59	71	24	24	18
\tilde{r}_t	38	47	80		42	53	47	53	53	47
\tilde{r}_t	0	17	50	17		8	8	17	8	0
<i>Early Group</i>										
\tilde{y}_t		27	21	25	0	7	7	27	20	7
$\tilde{\pi}_t$	80		87	75	67	69	63	81	81	81
\tilde{u}_t	0	20		50	0	60	67	13	20	13
\tilde{r}_t	75	75	75		67	25	25	25	25	25
\tilde{r}_t	0	33	67	33		0	33	33	33	0
<i>Later Group</i>										
\tilde{y}_t		59	100	89	67	68	95	68	91	68
$\tilde{\pi}_t$	9		50	0	0	0	0	0	0	5
\tilde{u}_t	50	100		0	0	50	100	100	50	50
\tilde{r}_t	22	33	100		33	64	55	64	64	55
\tilde{r}_t	0	11	0	11		11	0	11	0	0

Note: Entries are the percentage of forecasters for whom the column variable leads to a rejection of efficiency. Sample period varies across forecasters with a minimum of 10 observations. Before 1984 and after 1983 indicate when the forecaster data begin. Early Group contains forecasters whose forecasts begin before 1984 (mostly SPF) and the Later Group contains forecasters whose forecasts begin after 1983 (mostly WSJ).

diagonal) and the other errors exhibit even less. Only one cross-correlation is significant for half or more of the individual forecasters—the correlation between the unemployment error and the lagged inflation error. The rest are generally significant for one-third or fewer individual forecasters. However, for each error except the real interest rate, at least one lagged error is significant for one-third or more of the individual

Table 8

Regression Tests of Correlation for SPF and WSJ Individual Forecast Errors

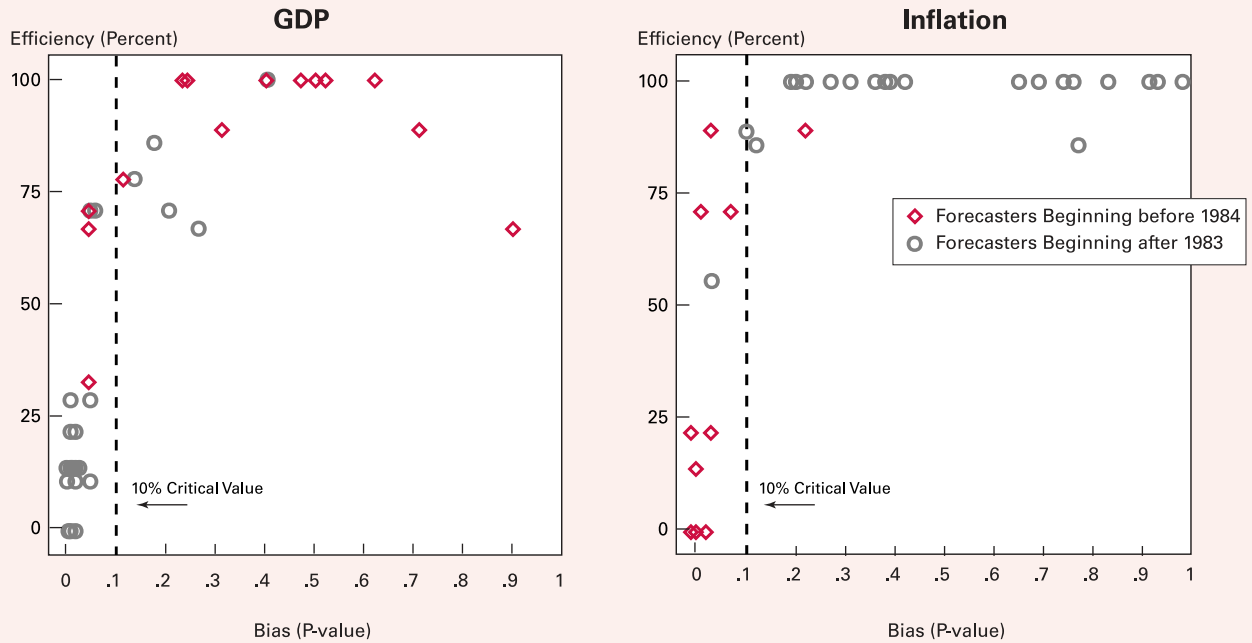
Percent Significant Correlation

Independent Variables	Forecast Errors				
	$\tilde{\epsilon}_t^y$	$\tilde{\epsilon}_t^\pi$	$\tilde{\epsilon}_t^u$	$\tilde{\epsilon}_t^i$	$\tilde{\epsilon}_t^r$
$\tilde{\epsilon}_{t-1}^y$	14	21	35	20	17
$\tilde{\epsilon}_{t-1}^\pi$	14	32	53	7	0
$\tilde{\epsilon}_{t-1}^u$	16	18	12	40	17
$\tilde{\epsilon}_{t-1}^i$	36	18	0	0	0
$\tilde{\epsilon}_{t-1}^r$	0	18	25	9	8

Note: The numbers represent the percentage of forecasters who have significant coefficient estimates in regressions of forecast errors on lagged errors.

Figure 4

Performance of Individual SPF and WSJ Forecasters



forecasters. To summarize, correlation in individual forecast errors is not rampant, but it is a problem for a significant fraction of individual forecast errors.

Evaluating Individual Performances

The tables summarizing the individual forecast tests conceal the relative performances of individual forecasters and thus preclude any assessment of whether some forecasters might be “better” than others. One way to quickly assess relative forecast performance is to examine the bias and efficiency test results for each individual forecaster, as shown in Figure 4. The figure contains scatter plots of bias and efficiency by forecaster for GDP and for inflation. The bias measure is the p-value, so larger numbers are better (p-values of less than 0.10 indicate significant bias). The efficiency measure is the percentage of variables in Z_{it} (out of a maximum of nine) for which efficiency is *not* rejected, so again larger numbers are better. The observations are separated by sample period around 1984.

Forecasters who have unbiased forecasts also tend to have efficient forecasts, in general. For both GDP and inflation, the figure reveals a positive, though nonlinear, relationship between the measures of unbi-

asedness and efficiency. Two distinct groups emerge. Forecasters with less than 50 percent efficiency have unequivocally biased forecasts, whereas most forecasters with at least 75 percent efficiency have unbiased forecasts. This sharp distinction among forecasters—biased and inefficient versus unbiased and efficient—contrasts with the result from the average forecasts, which were generally unbiased but inefficient. For the most part, the sample period is again the key distinguishing feature in the results, although more so for inflation than GDP.

This simple first pass at evaluating forecasters suggests important quality differences and motivates a more thorough investigation in the future. Such an investigation would benefit from an expanded data base that includes more forecasters whose data cross over subsample periods, as well as more forecasters within each subsample.

V. Summary and Conclusions

GDP and unemployment rate forecasts and, to a lesser extent, inflation forecasts veered off track in the second half of the 1990s. Although the errors are not

unusually large in historical perspective, they are economically significant and troubling—particularly from the perspective of monetary policymakers who require accurate forecasts to set interest rates appropriately. On average, macroeconomic forecasts are approximately unbiased, but they are inefficient and the forecast errors are characterized by improper correlation. These factors indicate that macroeconomic forecasts leave considerable room for improvement.

At least with regard to the period 1996 to 2000, no individual forecasters in the SPF or WSJ predicted macroeconomic conditions accurately and consistently. However, the brief and preliminary investigation of individual forecaster performance in this study provides evidence of differential abilities among forecasters. Much more data and analysis are required in this area before any firm conclusions can be drawn about the best forecasters.

Data Appendix

Survey of Professional Forecasters—The SPF has been conducted quarterly by the Federal Reserve Bank of Philadelphia since 1992 and formerly was conducted by the American Statistical Association (ASA) and the National Bureau of Economic Research (NBER). The SPF contains quarterly and annual forecasts of 24 economic variables, some of which extend back to 1968:Q4. The documentation and historical data are located at <http://www.phil.frb.org/econ/spf> and Stark (1997) provides additional details. Annual forecast data come from the November survey each year. See Table A1 for data details.

The Wall Street Journal—The WSJ publishes a survey of forecasters in early January and July each year. Currently, the WSJ survey contains nearly a dozen variables, but not all variables are available in every year and one-year-ahead forecasts also are not available every year (six-month-ahead forecasts are available instead). The annual data come from the January WSJ publication. Stephen McNees (1992) originally constructed this data base for the Federal Reserve Bank of Boston with data beginning in the mid 1980s, and I have updated it through 2000. See Table A1 for data details.

Blue Chip Economic Indicators—The BC survey (newsletter) was founded by Robert J. Eggert and is published

monthly by Aspen Publishers, Inc. It contains quarterly and annual forecasts of 15 economic variables since the mid 1970s. These data can be obtained from <http://www.bluechippubs.com/menu.htm>. The annual data come from the December BC survey. See Table A1 for data details.

Table A1

Forecast Data Sources and Definitions

Survey	Variable	Sample	Definition	Measure
SPF	y	1969–2000	Real GNP before 1992 Real GDP since 1992	Q4/Q4 % change
	π		GNP deflator before 1982 Total CPI-U since 1982	Y/Y % change
	u		Total civilian unemployment rate	Q4 value
	i	1982–2000	3-month T-bill rate	Q4 value
WSJ	y	1986–2000	Real GNP before 1992 Real GDP since 1992	Q4/Q4 % change
	π	1986–1998	Total CPI-U	Dec/Dec % change
	u	1986–1993	Total civilian unemployment rate	December
	i	1984–1998	3-month T-bill rate	December
BC	y	1977–2000	Real GNP before 1993 Real GDP since 1993	Y/Y % change
	π		GNP deflator before 1982 Total CPI-U since 1982	Y/Y % change
	u		Total civilian unemployment rate	Annual average
	i		3-month commercial paper rate (before 1983) 3-month T-bill rate (since 1983)	Annual average

References

- Baghestani, Hamid and Amin M. Kianian. 1993. "On the Rationality of U.S. Macroeconomic Forecasts: Evidence from a Panel of Professional Forecasters." *Applied Economics*, 25(7), pp. 869–78.
- Bonham, Carl and Richard Cohen. 1995. "Testing the Rationality of Price Forecasts: Comment." *The American Economic Review*, 85(1), March, pp. 284–89.
- Croushore, Dean and Tom Stark. 2000a. "A Real-Time Data Set for Macroeconomists: Does Data Vintage Matter for Forecasting?" Federal Reserve Bank of Philadelphia Working Paper No. 00–6, June.
- _____. 2000b. "A Funny Thing Happened On the Way to the Data Bank: A Real-Time Data Set for Macroeconomists." Federal Reserve Bank of Philadelphia *Business Review*, September/October, pp. 15–27.
- Granger, C.W.J. and Paul Newbold. 1986. *Forecasting Economic Time Series*. San Diego, CA: Academic Press.
- Keane, Michael P. and David E. Runkle. 1990. "Testing the Rationality of Price Forecasts: New Evidence from Panel Data." *The American Economic Review*, 80(4), September, pp. 714–35.
- _____. 1995. "Testing the Rationality of Price Forecasts: Reply." *The American Economic Review*, 85(1), March, p. 290.
- Koenig, Evan F., Sheila Dolmas, and Jeremy Piger. 2000. "The Use and Abuse of 'Real-Time' Data in Economic Forecasting." Board of Governors of the Federal Reserve System, International Finance Discussion Paper No. 684, November.
- Lamont, Owen. 1995. "Macroeconomic Forecasts and Microeconomic Forecasters." National Bureau of Economic Research Working Paper No. 5284, October.
- Laster, David, Paul Bennett, and In Sun Geoum. 1999. "Rational Bias in Macroeconomic Forecasts." *Quarterly Journal of Economics*, 114(1), February, pp. 293–318.
- Lim, Terrence. 2001. "Rationality and Analysts' Forecast Bias." *The Journal of Finance*, 61(1), February, pp. 369–85.
- Loungani, Prakash. 2000. "How Accurate Are Private Sector Forecasts? Cross-Country Evidence from Consensus Forecasts of Output Growth." International Monetary Fund Working Paper, April.
- McConnell, Margaret and Gabriel Perez-Quiros. 2000. "Output Fluctuations in the United States: What Has Changed Since the Early 1980's?" *The American Economic Review* 90(5), December, pp. 1464–76.
- McNees, Stephen K. 1992. "How Large Are Economic Forecast Errors?" *New England Economic Review*, July/August, pp. 26–42.
- _____. 1994. "Diversity, Uncertainty, and Accuracy of Inflation Forecasts." *New England Economic Review*, July/August, pp. 33–44.
- _____. 1995. "An Assessment of the 'Official' Economic Forecasts." *New England Economic Review*, July/August, pp. 13–23.
- Orphanides, Athanasios. 2000. "Activist Stabilization Policy and Inflation: The Taylor Rule in the 1970's." Board of Governors of the Federal Reserve System Finance and Economics Discussion Series No. 2000–13, February.
- Pindyck, Robert S. and Daniel L. Rubinfeld. 1976. *Econometric Models and Economic Forecasts*. New York: McGraw Hill.
- Robertson, John C. and Ellis Tallman. 1998. "Data Vintages and Measuring Forecast Model Performance." Federal Reserve Bank of Atlanta *Economic Review*, Fourth Quarter, pp. 4–20.
- Stark, Tom. 1997. "Macroeconomic Forecasts and Microeconomic Forecasters in the *Survey of Professional Forecasters*." Federal Reserve Bank of Philadelphia Working Paper No. 97–10, August.
- Theil, Henri. 1971. *Principles of Econometrics*. New York: John Wiley & Sons, Inc.
- Zarnowitz, Victor. 1985. "Rational Expectations and Macroeconomic Forecasts." *Journal of Business and Economic Statistics*, 3(4), October, pp. 293–311.
- Zarnowitz, Victor and Phillip Braun. 1993. "Twenty-two Years of the NBER ASA Quarterly Economic Outlook Surveys: Aspects and Comparisons of Forecasting Performance." In James H. Stock and Mark W. Watson, eds., *Business Cycles, Indicators, and Forecasting*, pp. 11–93. Chicago: The University of Chicago Press.
- Zarnowitz, Victor and Louis A. Lambros. 1987. "Consensus and Uncertainty in Economic Prediction." *Journal of Political Economy*, 95(3), pp. 591–621.