Research Program on Forecasting

Perspectives on Evaluating Macroeconomic Forecasts

By H. O. Stekler, George Washington University

RPF Working Paper No. 2010-002
http://www.gwu.edu/~forcpgm/2010-002.pdf

March 4, 2010

RESEARCH PROGRAM ON FORECASTING
Center of Economic Research
Department of Economics
The George Washington University
Washington, DC 20052
http://www.gwu.edu/~forcpgm

PERSPECTIVES ON EVALUATING MACROECONOMIC FORECASTS

H. O. Stekler, George Washington University

Over the past 50 or so years, I have been concerned with the quality of economic forecasts and have written both about the procedures for evaluating these predictions and the results that were obtained from these evaluations. In this paper I provide some perspectives on the issues involved in judging the quality of these forecasts. These include the reasons for evaluating forecasts, the questions that have been asked in these evaluations, the statistical tools that have been used, and the generally accepted results. (I do also present some new material that has not yet been published.) I do this in two parts: first focusing on short-run GDP and inflation predictions and then turning to labor market forecasts.

The process of forecasting involves a number of sequential steps. Part of that process is concerned with evaluating either the forecasts themselves or the methods that generated the predictions. This evaluation may occur when past forecasts are examined prior to preparing the next one or in a post mortem session to determine what went wrong and what can be learned from the errors. However, there are different perspectives or approaches for conducting these examinations. These differences may occur because some forecasts are model based while others were derived primarily from the judgmental approach. Originally, the evaluations were concerned with judging a particular model or individual. A more recent development has been to determine the value that the forecast has for the users of that prediction.

The original approach calculated a variety of statistics that measured the errors of the forecasts and then compared these errors with those generated by alternative methods or individuals. The newer approach for forecast evaluation is to base it on the loss functions of the

users (Pesaran and Skouras,2002). Elliott and Timmermann (2008) in summarizing the *theoretical* literature on how to evaluate forecasts take the same approach. These studies definitely suggest that the preferred evaluation methodology be based on decision-based methods, Pesaran and Skouras, however, noted that it has had limited use, and that most studies focused on statistical measures to evaluate the skills of the forecaster or accuracy of the model.[1] There were many reasons why the decision-based methods have not been used, including technical difficulties and the huge amount of data required to describe the decision environment, particularly the loss functions of the users.

The theoretical procedures provide the guidelines for undertaking forecast evaluations but the problems in applying them have not yet yielded much information about the quality of the forecasts or an understanding of the types of errors that occur or their causes. Even though we cannot evaluate the cost of the forecast errors in the context of decision functions, I will present one particular result where it was possible to use this approach.[2]

Statistical measures have been the most common method for evaluating forecasts, and they have provided many insights about the quality of the forecasts and their limitations. I will, therefore, focus on that approach. These measures also provide us with the ability to obtain information about the forecasting process and why particular errors occurred.

This paper proceeds as follows. I first present an old list of questions that should be addressed in any evaluation of macroeconomic forecasts. (Stekler, 1991). That paper also presented the statistical methods that could be used to address those questions. In the intervening

[1] West (2006) has still another view about forecast evaluation. He argues that these evaluations provide inferences about the characteristics of models. Thus the focus is exclusively on forecasts generated by models whether in sample or out of sample.

[2] In that regard, the aforementioned econometric procedures for conducting evaluations provided a rigorous theoretical methodology for the statistical measures.

20 years, forecasters have both developed new techniques for answering the original questions and asked additional questions. I will discuss these in the context of the original questions.

Our macroeconomic forecast evaluations have primarily been concerned with the predictions of GDP growth and inflation. I will, therefore, summarize some of the findings relating to these two variables. In making macro forecasts, economists also estimate the unemployment rate, but these forecasts have not been analyzed as extensively.

I will then examine the literature that analyzes these labor market forecasts and will present the latest findings. Moreover, in analyzing labor markets in a macroeconomic growth context, projections of annual employment by industry and occupation are also issued. Not as much attention has been paid to the procedures for evaluating these projections. I will present some findings about these forecasts, utilizing a statistic that has not conventionally been used in evaluations. This method can also be applied to the population projections of the Census Bureau. I conclude with a summary of the findings and suggest topics that warrant further research.

I. Questions Addressed in Macroeconomic Forecast Evaluations

In 1991, I asked a number of questions, some of which are related, that can be and have been used in forecast evaluations. (Stekler, 1991) These questions are all statistical in nature and describe the characteristics of the forecasting method or the particular forecaster. The main questions are (1) How good is Method A (Forecaster X)? (2) Do the forecasts show systematic errors? (3) Are all forecasters (methods) equally good? (4) Is Method A (Forecaster X) significantly better than Method B (Forecaster Y)? (5) Does Forecast M contain information not in Forecast N? (6) Does Forecaster X produce forecasts that are useful to users?

A. How good is Method A (Forecaster X)?

If the forecasts are quantitative, the first question is answered by using some error metric, usually mean square error or mean square percentage error,[3] and comparing it with the similar error metric of a benchmark or naïve model. On the other hand, different procedures are used to assess non-quantitative macroeconomic forecasting techniques, which are primarily concerned with predicting whether a cyclical turn will occur. These forecasting methods are based on indicators. Originally rules were used for determining whether an indicator was signaling a turn. (See Stekler, 1989). More recently, models that predict the probability of a cyclical turn have been developed, and probability scoring rules such as the Brier Score have been used to evaluate these forecasts. SeeXXXXXXXXXX

B. Bias and Efficiency

Accurate forecasts should be unbiased (not have systematic errors) and should use all available information. Bias and efficiency tests are used to determine whether there are systematic errors. These tests are usually derived from the Mincer-Zarnowitz (1969) equation:

$$A_t = \alpha + \beta F_t + e_t, \quad \text{where} \qquad (1)$$

A_t and F_t refer to the actual and forecast values. The condition for unbiasedness and weak form efficiency is that $\alpha = 0$ *and* $\beta = 1$. An alternative test for bias is:

$$A_t - \beta F_t = c + e_t, \qquad (1a)$$

with the null that $c = 0$. (Holden and Peel, 1990). These tests are applied to one individual's forecasts at one horizon.

A newer and more sophisticated methodology has been developed by Davies and Lahiri, (1995, 1998), and has been applied to surveys that contain the forecasts of many individuals or organizations, e.g. the Survey of Professional Forecasters (SPF) or Blue Chip Forecasters. The

[3] The mean square error criterion is associated with a quadratic loss function. (See Elliott and Timmermann, 2008). Another metric is mean absolute error.

use of this methodology permitted an analysis of multi-dimensional forecasts, i.e. many forecasters each making a prediction for a target year at several horizons.[4] The errors can be decomposed as (2a and 2b)

$$A_t - F_{ith} = \Phi_i + \lambda_{th} + e_{ith} \qquad (2a)$$

$$\lambda_{th} = \sum u_{tj} \quad \text{where} \qquad (2b)$$

F_{ith} is the forecast made by the ith forecaster for year t at a h month horizon; Φ_i represents the specific bias of each individual; λ_{th} represents the shocks that were not anticipated. This model can identify the specific sources of each forecaster's errors.

Using the Davies-Lahiri methodology it is no longer necessary to confine the evaluation of the predictions obtained from a survey to the "consensus" forecast. It has the inherent advantage that the opposing biases of individuals can make the mean (median) forecast look unbiased even when, in fact, the individual forecasts from which the mean is calculated are all biased.

C. Comparing Forecasters and Benchmarks

Non-parametric procedures have been used to determine whether all forecasters are equally accurate. For each set of forecasts, the errors of each forecaster are ranked according to their accuracy. If the accuracy of all forecasters were equal, their ranks would have the same expected values. It is, thus possible to test the hypothesis that all forecasters have equal ranks (and are equally good). The chi-square goodness-of-fit test statistic, X^2, is used. (See Batchelor, 1990).

[4] This method can also be applied to one forecaster making several forecasts for the same horizon. (See Clements et al. 2007).

The fourth question asks whether a particular method (forecaster) is significantly more accurate than another method (forecaster). Originally, Theil's U coefficient was the basis of comparison:

$$U = \sqrt{(e_{f,t})^2} / \sqrt{(e_{n,t})^2}\ , \qquad (3)$$

where $(e_{f,t})$ is the error of the forecast that is being evaluated and $(e_{n,t})$ is the error of the naïve benchmark. This naïve model can be either a no-change or the same change as last period prediction. At a minimum, the forecasts should be more accurate than naïve models and U must be less than 1. Statistical models, such as ARIMA, have also been used as benchmarks, and new statistics for comparing models have been developed.

Currently the Diebold-Mariano (DM) statistic (1995) is the preferred methodology for testing whether there is a statistically significant difference in accuracy between any two sets of forecasts. That statistic (Eq 4) has been modified by Harvey et al. (1997) which results in an improvement in the behavior of the test statistic for moderately-sized samples:

$$S_1^* = S_1\left(\frac{T+1-2(h+1)+h(h+1)/T}{T}\right)^{-\frac{1}{2}},\ S_1 = \frac{\bar{d}}{[\hat{V}(\bar{d})]^{1/2}} \qquad (4)$$

where h is the horizon, $\bar{d}$ is the mean absolute difference of the prediction errors, $\hat{V}(\bar{d})$ is the estimated variance, S_1 is the original DM statistic, and S_1* is the modified DM statistic. The modified Diebold-Mariano test statistic is estimated with Newey-West (1987) corrected standard errors that allow for heteroskedastic autocorrelated errors.

Several procedures have been developed to determine whether one forecast contains information not embodied in another procedure. One involves combining the two sets of forecasts. If the variance of the *combined* forecasts is not significantly less than the variance of the prediction which is being analyzed, then this particular forecast does not contain additional

useful information. This analysis is similar to the Hendry-Richard (1983) concept of encompassing, where a model that encompasses another contains the information of the latter.

D. Directional Accuracy: A New Approach

Even though the analysis does not directly use utility functions, the final question listed above relates to the usefulness of a forecast to a decision maker. It concerns the directional accuracy of the forecast. Merton (1981) in analyzing financial forecasts indicated that they had value if the signs of the predicted and actual changes were similar. The various tests that have been implemented seek to determine whether the sign of the forecast change is probabilistically independent of the actual change. If the hypothesis that the forecasts are independent of the observed events is rejected, then the forecasts can be said to have value.

Schnader and Stekler (1990) provided another interpretation of this test. We argued that testing whether the forecasts have value is the same as determining whether (in the sense of predicting the direction of change) the forecast differed significantly from a naïve model which continuously predicted up (down). The profession now categorizes the various tests as measuring directional accuracy. This concept can be illustrated either for the GDP or inflation predictions that are made separately or when both variables are examined together.

1. One Variable

Most macroeconomic forecast evaluations focus on GDP and inflation. Basically, when the real GDP and inflation forecasts are each evaluated separately, they are grouped into two categories. The GNP/GDP forecasts are categorized according to whether GDP growth was positive or negative, and the inflation categories depend on whether inflation increased or decreased.[5] A 2x2 contingency table is created that compares the predicted outcome of a

[5] No change is classified with the negative changes. Note that we are focusing on the direction of change in the inflation rate, which is equivalent to measuring accelerations and decelerations of the price level.

variable with the actual outcome of that variable (Table 1). [6]

Table 1: The Relationship between Predicted and Actual Outcomes			
Predicted Outcome	Actual Outcome		
	> 0	≤ 0	
$\Delta Y > 0$	n1	N2-n2	n
≤ 0	N1-n1	n2	N-n
	N1	N2	N

For notation we have a total of *N* observations where for *n1* of them both the actual and the predicted are positive and for *n2* of them both the actual and the predicted are negative. We have *n* observations where the predicted outcome is positive and *N-n* observations where the predicted outcome is negative (or zero). We also have *N1* observations where the actual outcome is positive and *N2 = N – N1* observations where the actual outcome is negative (or zero). The Pesaran-Timmerman (1992) statistic along with the chi-square and Fisher's exact test can be used to test this hypothesis.[7]

The Pesaran-Timmermann statistic for predictive performance for an *m* x *m* contingency table with a total of *N* observations is:

$$s_n = \sqrt{N}\mathbf{V}_n^{-1/2} S_n$$

where we have the following:

$$\hat{\mathbf{V}}_n = \left(\frac{\partial f(\mathbf{P})}{\partial \mathbf{P}}\right)_{\mathbf{P}=\hat{\mathbf{P}}} \hat{\mathbf{\Omega}} \left(\frac{\partial f(\mathbf{P})}{\partial \mathbf{P}}\right)_{\mathbf{P}=\hat{\mathbf{P}}}.$$

$\hat{P}_{ij} = n_{ij} / N$, where n_{ij} is the number of observations in the *ij* category of the contingency table.

$\hat{\mathbf{\Omega}} = \hat{\mathbf{\Psi}} - \hat{\mathbf{P}}\hat{\mathbf{P}}'$, where $\hat{\mathbf{\Psi}}$ is an m^2 x m^2 diagonal matrix with the elements of $\hat{\mathbf{P}}$ on the diagonal.

[6] The contingency table methodology is used to test whether the sign of the predicted change is probabilistically independent of the sign of the actual change. This is also a test of the hypothesis that the forecasts are more accurate than those of a naïve random walk model in predicting the direction of change (see Stekler, 1994, page 497).
[7] The Chi-square and Fisher's exact test are well known and are not presented here.

$S_n = \sum_{i=1}^{m} \left(\hat{P}_{ii} - \hat{P}_{i0}\hat{P}_{0i} \right)$, where $\hat{P}_{i0} = n_{i0} / N$. $\hat{P}_{0i} = n_{0i} / N$, where n_{i0} and n_{0i} represent the i[th] row and column totals respectively.

Pesaran and Timmermann (1992) present their results based on the square of this test statistic in order to more easily compare it to the Chi-square goodness of fit statistic. This test statistic has a Chi-square distribution with one degree of freedom.

2. Several Variables

All of the questions that were discussed above were concerned with evaluating the forecasts of one variable at a time. However, in preparing a particular macroeconomic forecast, individuals are concerned with the outlook for both the growth rate and the rate of inflation. The accuracy of this overall forecast thus depends on how well both variables are predicted simultaneously. Thus, it is necessary to use a different contingency table for evaluating the directional accuracy of these macroeconomic forecasts. Sinclair et al. (forthcoming) showed that the simultaneous directional accuracy of the two variables can be evaluated by using a 4x4 contingency table[8] rather than the 2x2 table that had been used in assessing each variable individually.

In the expanded 4x4 table, instead of simply categorizing based on the separate GDP growth or inflation predictions, forecasts about the state of the economy are grouped into four categories: (1) GDP growth positive, inflation increasing, (2) GDP growth positive, inflation decreasing, (3) GDP growth negative, inflation increasing, and (4) GDP growth negative, inflation decreasing. The statistical tests are generalized versions of those used when the

[8] Naik and Leuthold (1986) also used a 4x4 contingency table in their qualitative analysis of forecasting performance. Their study focused on a different topic-the ability to predict turning points. (Also see Kaylen and Brandt, 1988).

forecasts were analyzed separately.[9] Table 2 illustrates a 4x4 contingency table when the directional accuracy of the GDP growth and inflation forecasts of the Federal Reserve are evaluated jointly.

Table 2: The 4x4 Contingency Table for the Zero Month Lead

	Actual Outcome			
	$\Delta GDP > 0$, $\Delta inf > 0$	$\Delta GDP > 0$, $\Delta inf \leq 0$	$\Delta GDP \leq 0$, $\Delta inf > 0$	$\Delta GDP \leq 0$, $\Delta inf \leq 0$
Predicted Outcome				
$\Delta GDP > 0$, $\Delta inf > 0$	49	13	1	1
$\Delta GDP > 0$, $\Delta inf \leq 0$	7	43	0	4
$\Delta GDP \leq 0$, $\Delta inf > 0$	1	2	4	2
$\Delta GDP \leq 0$, $\Delta inf \leq 0$	0	3	5	4

Source: Sinclair et al. (forthcoming)

3. Test Statistics

The statistical methodology tests whether or not the forecasts predict the associated directions of change. There are at least three test statistics that can be used to test the hypothesis that the forecasts fail to predict the observed events.[10] Two test statistics focus on independence. These test statistics are the Chi-square and Fisher's exact test. The Pesaran-Timmermann (1992) statistic specifically focuses on predictive failure. The forecasts are said to have value only if the null hypothesis of predictive failure is rejected. Pesaran and Timmermann's predictive-failure test is particularly useful in the case where we undertake a joint evaluation of GDP growth and inflation forecasts. Their test does not require that the two forecasts be independent of each

[9] There is, however, a difference in interpretation once we go beyond the simple 2x2 case. In particular, the 2x2 contingency table tests for predictive failure of only one variable whereas the 4x4 contingency table tests for predictive failure of both variables. Moreover, in the 2x2 case, the hypothesis of predictive failure is equivalent to the hypothesis that the actual and predicted values of the variable are independent of each other. As discussed in Pesaran and Timmermann (1992), however, for the 4x4 case they are no longer equivalent. For our contingency table, independence implies predictive failure, but not vice versa.

[10] Merton (1981) and Henriksson and Merton (1981) had used a test based on the hypergeometric distribution. This is identical to Fisher's exact test. Their test assumes known row and column frequencies, which is not assumed for the Pesaran-Timmermann test.

other. Since output and inflation may be predicted from the same forecasting model, this is an important consideration. In this particular case, the probability of the pattern of these forecasts occurring by chance is less than .001. Sinclair et al. (forthcoming) thus concluded that the Fed forecasts for the current quarter yielded an accurate view of the state of the economy.

E. Policy Forecast Errors: An Example

In general to obtain a quantitative measure of the economic costs of forecast errors, the decision rule of the user of the prediction must be known. However, there is at least one case where the cost of forecast errors can be measured without knowing an explicit decision rule because there is another criterion for evaluating forecasts: Are financial market and betting market decisions based on those forecasts profitable? (Leitch and Tanner, 1991).

Macroeconomic forecasts, however, cannot be evaluated in this way, because there is no generally acceptable way of calculating their value to policy makers. There is an exception if the rule guiding policy decisions is known. Sinclair et al. (2009) showed that it was possible to evaluate the quantitative forecasts of the Federal Reserve within the context of the Taylor Rule which is assumed to be the one that guides the Fed in setting monetary policy.[11] The assumption was that the Fed's forecasts of multiple series were generally generated for a specific policy purpose, as inputs for monetary policy. In this case, an assessment of the quality of the quantitative forecasts of two or more variables depends on the relative importance of each to the Fed.

Specifically, let $P^{f}_{t,t+h}$ be a policy decision at time *t* that is a linear function of the h-step ahead forecasts of $N \geq 1$ variables ($x^{f}_{i,t+h}, i = 1,...N$). The superscript *f* indicates that the policy decision is based on forecasts rather than the actual outcomes of the variables:

[11] The literature assumes that the Taylor rule is an approximation to the decision rule of the Fed. Also note that that this procedure is in the framework of a decision-based forecast evaluation discussed above.

$$P^f_{t,t+h} = p(x^f_{1,t+h},...x^f_{N,t+h}). \quad (5)$$

If policymakers have perfect foresight, the policy decision would simply be P_t without the superscript *f*:

$$P_{t,t+h} = p(x_{1,t+h},...x_{N,t+h}). \quad (6)$$

However, because policy is based on forecasts, rather than on the actual data, policy is subject to errors which are functions of the mistakes made in forecasting the underlying variables $x_{i,t}, i = 1,...N$. The difference between the actual policy and the policy that would have been pursued under perfect foresight is called the policy forecast error (PFE):

$$PFE_t = P_{t,t+h} - P^f_{t,t+h} = p(x_{1,t+h},...x_{N,t+h}) - p(x^f_{1,t+h},...x^f_{N,t+h}) = e(e_{1,t+h},...e_{N,t+h}), \quad (7)$$

where the $e_{1,t+h},...e_{N,t+h}$, are the forecast errors associated with the individual series. Thus the PFE is composed of the individual forecast errors weighted by their importance in the policy rule. According to the forward-looking Taylor rule[12] , the Fed, sets a target federal funds rate, i^{Tf}_t , based on equation (8), where, as above, the superscript "f" denotes that the target is based on forecasted variables.[13] The Fed's policy decision $(P^f_{t,t+h})$ is written as:

$$P^f_{t,t+h} = i^{Tf}_t = r^* + \pi^f_{t+h} + 0.5(\pi^f_{t+h} - \pi^*) + 0.5(y^f_{t+h} - y^*), \quad (8)$$

where r* is the equilibrium real interest rate, π^* is the Fed's implicit inflation rate target, and y* is the potential output growth rate.[14]

---

[12] Although Taylor (1993) originally proposed his rule as an empirical description of past Fed policy actions, Woodford (2001a, 2001b) has shown that the Taylor rule can also be justified based on a firm theoretical foundation.

[13] Following Orphanides (2001), we assume that the Fed uses the Greenbook forecasts in their decision rule. The members of the FOMC also make their own forecasts, but have access to the staff forecasts of the Greenbook when doing so. For an evaluation of those forecasts, see Romer and Romer (2008).

[14] While the output gap is typically used in the Taylor rule, the growth rate is typically used in forecast evaluation. The growth rate of the actuals is approximately $\ln(Y_t) - \ln(Y_{t-1})$, whereas the growth rate of the forecasts is approximately $\ln(Y^f_t) - \ln(Y_{t-1})$. Thus, when we subtract one from the other for the policy forecast error, we have

The actual outcome in period $t+h$, however, may differ from the Fed's forecasts. Therefore, if the members of the FOMC had known the actual values for π_{t+h} and y_{t+h} (i.e. if they had perfect forecasts or perfect foresight), they would have chosen a (potentially different) federal funds rate. Consequently, their policy decision under perfect foresight ($P_{t,t+h}$) would have been:

$$P_{t,t+h} = i_t^T = r^* + \pi_{t+h}^A + 0.5(\pi_{t+h}^A - \pi^*) + 0.5(y_{t+h}^A - y^*), \quad (9)$$

where π_{t+h}^A and y_{t+h}^A represent the actual realizations of π_{t+h} and y_{t+h}. The difference between i^{Tf} and i^T measures the difference in the Fed funds rate that occurs because of inaccurate forecasts of output growth and inflation and thus represents the Federal Reserve's policy forecast error, PFE_t:

$$PFE_t = i_t^T - i_t^{Tf} = 1.5\left(\pi_{t+h}^A - \pi_{t+h}^f\right) + 0.5\left(y_{t+h}^A - y_{t+h}^f\right). \quad (10)$$

The differences, $\left(\pi_{t+h}^A - \pi_{t+h}^f\right)$ and $\left(y_{t+h}^A - y_{t+h}^f\right)$, are the Fed's forecast errors for the inflation rate and real output growth respectively. Given the PFEs, the evaluation procedures are similar to those used in judging individual forecast errors.

Using this methodology, Sinclair et al. (2009) were able to evaluate the impact that forecast errors had on the Fed's monetary policy as characterized by the Taylor rule. They found that the Fed's policy forecast error was in general unbiased and significantly smaller than the errors that would have resulted from naïve forecasts but not always from the SPF predictions. Nevertheless, the mean absolute policy forecast error of the Fed forecasts was approximately 1% (100 basis points).

$\ln(Y_t) - \ln(Y^f_t)$. Approximating the output gaps in the same manner, we have $\ln(Y_t) - \ln(Y^*)$ and $\ln(Y^f_t) - \ln(Y^*)$, so again we have $\ln(Y_t) - \ln(Y^f_t)$. It is this result that permits us to use the growth rate in order to construct the PFEs. This analysis does assume, however, that potential output, Y*, is known rather than a forecast. This assumption is based on the lack of forecasts for this variable in the Greenbook. For a discussion of the role of real time output gap estimates and the Taylor rule, see Orphanides (2001).

II. Findings from Forecast Evaluations (of GDP growth and Inflation)

There have been many studies that have reported on the accuracy of the forecasts of the growth of GDP and inflation. I have been involved in two survey papers that have summarized and synthesized the results of these studies. One looks at US and UK forecasts (Fildes and Stekler, 2002). The other does a similar analysis of the G7 (excluding the US) predictions. (Stekler, 2008). By comparing the forecasts of various countries, we can determine whether the findings are robust. The focus will be on five topics (1) directional errors, (2) biases and systematic errors, (3) the magnitude of the errors, (4) the source of the errors, and (5) the trend, if any, in forecast accuracy.

A. Directional Errors

There are very few analyses about directional errors because most forecast evaluations focus on the magnitude of the quantitative errors. However, Fildes and Stekler (2002) note that most US and UK recessions were not predicted in advance, but neither did economists make many predictions of peaks that did not occur.[15] More recently, Sinclair et al. (forthcoming) analyzed the directional accuracy of the Fed's forecasts of GDP and inflation and showed that the predictions of increases and decreases in the inflation rate were not associated with the actual changes in that rate. When, however, the directional accuracy of the GDP and inflation predictions were analyzed jointly, on average the Fed's forecasts for the current quarter and one quarter-ahead period yielded an accurate view of the state of the economy.

The record for other countries is no better. The turning points in German GDP were not predicted, but the accelerations and decelerations of the growth rate were forecast accurately. (Stekler, 2008, summarizes the literature relating to the forecasts of the G7 countries and

[15] The US false turns predicted in 1978-79 were an exception, but real GNP did decline for two quarters during this period.

indicates that the results apply equally to private forecasters, research institutes and international organizations). The evidence suggests that forecasters are not able to predict turning points in advance and may even have difficulty in detecting them quickly once they have occurred.

B. Biases and Systematic Errors

Most evaluations examine the rationality and efficiency of the predictions in order to determine whether they could have been improved. Stekler (2001) reviewed a large number of studies and concluded that there was no definitive evidence that the US inflation forecasts displayed weak form informational efficiency. While more of the US growth forecasts did not reject the null of informational efficiency, these results were also mixed.[16] The results were dependent on the database that was examined, the years that were examined, and the methodology that was employed. However, most of these analyses did not test whether the forecasts were truly inefficient or whether the errors could be attributable to asymmetric loss functions.

Forecasters also made systematic errors. They overestimated the rate of growth during slowdowns and underestimated it during recoveries and booms. Similarly, inflation was underpredicted when it was rising and overpredicted when it was declining. (See the surveys of Fildes and Stekler (2002) and Stekler (2008) for the specific studies from which these results were obtained). Fildes and Stekler concluded: *"these errors occurred when the economy was subject to major perturbations, just the time when accurate forecasts were most needed."* (p.442).

C. Magnitude of the Errors

Although these qualitative findings about directional and systematic errors are important to our understanding the forecasting process, most evaluations have also provided quantitative estimates of these errors. Fildes and Stekler (2002) reported that the Mean Absolute Error of

[16] The UK forecasts yielded similar results.

annual US and UK GDP growth forecasts was around 1%. Newer studies found similar results for the G7 countries, but Oller and Barot (2000) had found that the errors were larger for some of the other European countries. (Stekler, 2008).

When quarterly GDP estimates were examined, the previous papers did not all use identical procedures for calculating the MAEs.[17] Consequently, our findings are not as complete. We do know that there is a substantial improvement in accuracy when the forecasting task switches from predicting what will happen in the next quarter to estimating the level of activity in the current quarter. This is largely attributable to the availability of actual data for the current period.

The inflation forecasts seemed to have improved. The earlier US inflation forecasts had MAEs between 1.0 and 1.4%, but Stekler's (2008) survey of G7 forecasts showed that those errors were now between 0.5% and 0.75%. The reduction may be attributable to the lower inflation rates that have been observed in the past several decades.

D. Have the Forecasts Improved?

Given the number of papers that have evaluated macroeconomic forecasts, it is surprising how few have asked whether the quality of the predictions has improved over time. The problem is not the lack of data, for we have 40 years worth of forecasts for some countries. The findings of those studies that have examined this issue are contradictory, and thus there are no definitive conclusions. For example, Heilemann and Stekler (2003) examined German forecasts and adjusted the errors for the difficulties in predicting the relevant periods, but the results were mixed. Dopke and Fritsche (2006) also looked at German forecasts and suggested that accuracy may have improved. As for the predictions of international organizations, Vogel (2007) showed that the accuracy of the OECD forecasts had improved, but Timmerman (2007) indicated that the

[17]Some authors transform the errors into annual growth rates; others do not.

quality of the IMF forecasts had deteriorated. These findings are consistent with those summarized by Fildes and Stekler (2002). We conclude that despite all the resources that have been devoted to forecasting, there is no clear evidence that accuracy has improved.

This finding suggests that we may have reached the limits of forecastability. Heilemann and Stekler (2003) have investigated this hypothesis. We calculated the ex post forecast errors generated by simulations obtained from econometric models. These models can serve as a benchmark of the maximum accuracy that is attainable because they are free from errors caused by wrong assumptions about the predetermined variables and the inability to capture the dynamics of multiperiod forecasts. Heilemann and Stekler found that the model's inflation errors were very similar to those that were made ex ante for the same period. This result indicates that the accuracy of the inflation forecasts could not have been improved substantially. On the other hand, the model's growth rate errors were substantially smaller than the ex ante errors suggesting that the quality of the ex ante real sector forecasts can still be improved.

E. Recessions Sources of Error

A model based forecast can be decomposed into various sources. The forecast depends upon the econometric specification, the exogenous variables and the corresponding predictions of these variables, and any adjustment that the economist makes to the model output. There are analytical difficulties associated with determining why each of these errors occurred. One example of this difficulty occurs when econometricians make assumptions about the exogenous variables rather than model adjustments to subjectively influence their forecasts.

Nevertheless, there is agreement that recessions are a significant cause of some of these errors. A large portion of GDP forecast errors are attributable to the failure to predict the occurrence of recessions. If recessions and booms are caused by events such as changes in asset

prices, that it is assumed cannot be predicted, then the recessions are themselves unforecastable. Fair (2009) found that, ex post, some recessions could be predicted even if some key exogenous asset variables were estimated using only baseline paths.[18] The failure to adequately predict the other recessions could be explained by the inability to estimate some or all of these key exogenous variables.

E. What Have We Learned?[19]

The evidence about the macro forecasts that has been presented here is very robust. The findings of Fildes and Stekler that primarily related only to US and UK forecasters are similar to those relating to the G7 economists. Both studies found that recessions are not predicted in advance and account for a significant portion of the quantitative errors. Neither study was able to show that forecast accuracy had improved and both found that there were systematic errors. There may be a quantitative limit beyond which forecast accuracy cannot be improved. (Heilemann and Stekler, 2003).

Finally, we now have a somewhat better understanding of the forecasting process. We have learned that forecasts for a horizon longer than 12-18 months might not be valuable. (Vuchelen and Guitierrez, 2005b; Isiklar and Lahiri, 2007). We also know more the causes of bias. Batchelor (2007) showed how the systematic errors or "bias" were related to the forecasters' optimism (pessimism) and conservatism in revising their predictions. He notes that standard rationality tests are not appropriate if there has been a structural break. The pattern of the errors can then provide a way of understanding the forecasters' learning process about the

[18] The exogenous asset variables in the Fair model are equity prices and housing prices. His simulations also assumed that import prices and exports as well as the asset variables could only be estimated using baseline benchmarks.

[19] This summary refers only to the findings referred to in the text. There are many other topics in the forecasting literature that were not reviewed and are beyond the scope of this paper. These include the quality of the data (Oller and Teterukovsky, 2006); leading indicators (Allen and Morzuch, 2006); the role of judgmental forecasting (Lawrence et al., 2006). In addition one could investigate the value of combining forecasts, data revisions and which actuals to use in conducting an evaluation.

impact of this structural break. Isiklar and Lahiri (2007) use forecast revisions to explain the behavioral characteristics of forecasters, i.e. how they react to news and when is news important.[20] We know that there is much more work to be done in determining the importance of asymmetric losses, the sources of biases, the limits of accuracy, etc. Much can be learned if we undertake more studies, about the sources of error, similar to Heilemann's (2002) study.

III. Labor Market Forecasts

We now turn to a discussion of labor market forecasts. There are many fewer studies of these types of forecasts. Consequently, our results will be less informative. Before I discuss the results of evaluations of the forecasts of these variables, I want to briefly note that several labor market series are used as indicators or predictors about the overall state of the economy. Next I consider the short run quantitative forecasts of the US unemployment rate. Finally, I consider long run projections and the procedures for evaluating them. I also show that the methodology for evaluating long-run labor market forecasts can be used to analyze other types of long-run projections.

A. Labor Market Series as Indicators

The US unemployment rate moves counter cyclically and may display an asymmetric relationship with changes in GDP. It may have a short lead (or be coincident) at business cycle peaks but it lags at the troughs. (Montgomery et al., 1998). The unemployment rate is not considered either a leading or coincident indicator. There are, however, two other series that are considered leading indicators of cyclical movements and one series that is considered a

[20] In the forecasts made for year t, the most important revisions occur at the end of t-1.

coincident indicator. These are series that are included in the various indicator indexes currently compiled and published by the Conference Board.

The leading series are (1) average weekly hours in manufacturing and (2) average weekly claims for unemployment insurance.[21] While Diebold and Rudebusch, (1989; 1991) have evaluated the composite leading indicators, I have found only one paper that examined the forecasting behavior of either of those series. (Seip and McNown, 2007). Seip and McNown examined the behavior of average weekly hours, but their analysis produced contradictory findings. For example, they examined the Granger causality between movements in the weekly hours series and changes in the Federal Reserve Board Index of Industrial Production. Their results showed that the hours series is a lagging indicator with respect the Index of Production which is a coincident indicator. On the other hand, using sophisticated phase analysis, they found that the timing at turning points suggested that it was a leading (but not that accurate an) indicator.

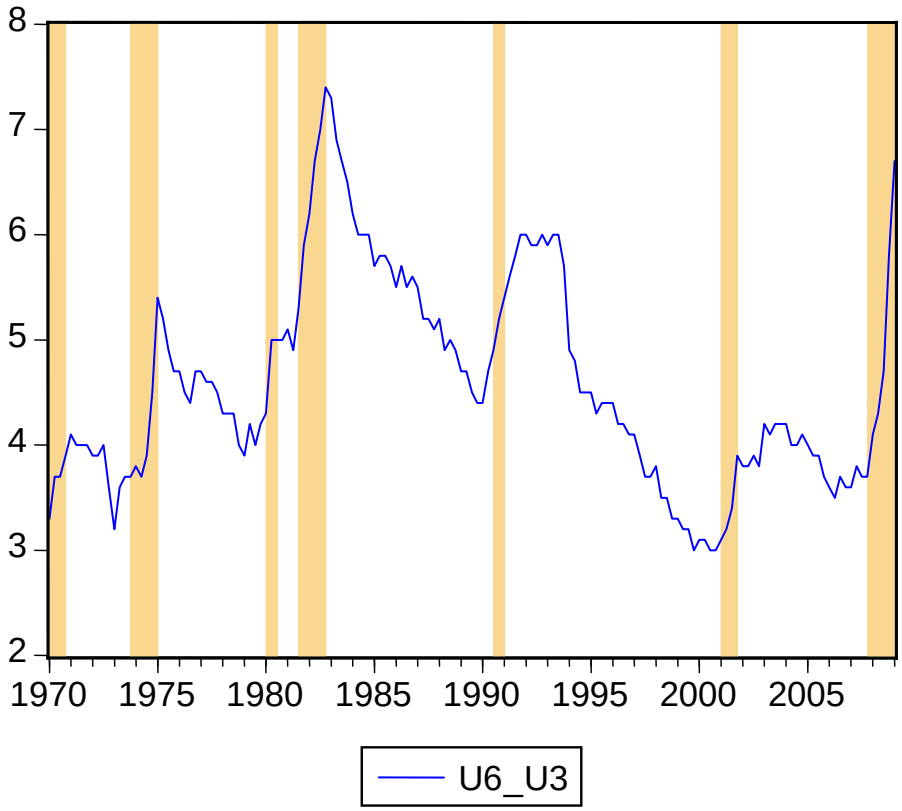
It is possible that there may be another labor market series that could be a leading indicator. The official US unemployment rate series is not the only measure of labor slack in the economy; the US Department of Labor also compiles other measures of unemployment. The official or conventional unemployment rate, as defined is called U3 and measures the total number of unemployed persons as a percent of the civilian labor force. The broadest BLS measure of unemployment is called U6. It includes two additional categories: (1) people who have left the labor force because they have become "discouraged" by failing to find employment and (2) individuals who are working part time but would prefer to be full-time employees

[21] The Conference Board has constructed a Composite Leading Index and these series are included in that Index. They had also been included in earlier composite indexes that had been constructed by the US Department of Commerce.

The difference between the two rates, therefore, represents the degree of labor underutilization that is not captured by the traditional unemployment rate. Figure 1 displays the difference between the U6 and U3 series that seems to indicate that this difference leads the turns of the US business cycles for the period 1970-2009. The NBER-dated recessions are shaded. However, there has been no rigorous and systematic evaluation of this series to determine whether, in fact, this series is an adequate leading indicator.

The Conference Board also compiles and publishes a composite index of coincident series. The movements in this index roughly track the cyclical movements of the economy. There are four series in this composite index, with "employees on non-agricultural payrolls" representing the labor market. We should note the importance of using labor market data for forecasting cyclical movements, but the procedures for evaluating these time series as indicators are beyond the scope of this paper.

Figure 1: Difference Between U3 and U6



Source: John Dougherty

B. Modeling and Forecasting the Unemployment Rate

Neftci (1984) found that the US unemployment rate had an interesting characteristic. It displayed asymmetric behavior because the probability of a decline following two previous declines differed from the probability of an increase given two prior increases. This suggested that a linear model would not adequately explain the behavior of this series. Subsequently, a number of univariate non-linear models were developed to explain and then forecast this series.[22] More recently, Milas and Rothman (2008) went one step further by developing non-linear *multivariate* models to explain the US unemployment rate.

[22] Skalin and Terasvirta (2002) provide a list of studies that have employed nonlinear models to estimating unemployment rates. Swanson and White (1997) select models on their ability to predict macroeconomic variables, including the unemployment rate, in real time. Clements and Krolzig (2003) survey the development of asymmetric business cycle models and develop statistical tests, but do not apply these models to US unemployment data.

Only a small number of these non-linear models have actually been used to forecast the US unemployment rate. Rothman (1998) used six models and, based on out of sample recursive simulations, concluded that their performance was better than that of a linear model.[23] Montgomery at al. (1998) undertook a more comprehensive evaluation. They used a Threshold Autoregressive (TAR) Model to generate simulated recursive forecasts. These forecasts displayed smaller errors than were generated by the linear ARIMA model. Moreover, although no formal statistical tests were used, the forecasts of the TAR model appeared to be unbiased. Similarly, neural network models (Moshiri and Brown, 2004) and non-linear non-parametric models (Golan and Perloff, 2004) were superior to linear models in forecasting the unemployment rate. In fact, Golan and Perloff indicated that their model was superior to the non-linear TAR model.

Unfortunately, it is unlikely that any of these models will become the standard methodology for forecasting the unemployment rate. The median SPF forecast was more accurate than either the TAR or the non-parametric non-linear models. In addition, Golan and Perloff found that the Michigan structural model also generated smaller errors.

Given these results, it is appropriate to also report some findings about non-model forecasts of the unemployment rate. The median SPF forecast was not only more accurate than the model predictions, but it was also superior to the Federal Reserve's Greenbook estimates for the period 1983-2004. (Baghestani, 2008). The SPF errors are asymmetric, with mean errors during expansions less than 0.1%, while during recessions, those errors sometimes exceed 1.0%. (Montgomery et al. 1998). The Greenbook estimates are unbiased, but the errors are also asymmetric. (Joutz and Stekler, 2008).

[23] Pool and Speight (2000) had a similar finding for the UK and Japanese economies.

There is additional information about the quality of non-model unemployment rate forecasts. Carroll (2003) noted that the SPF forecasters were more accurate than those obtained from the Michigan Household Surveys. The households eventually did update their forecasts to conform to those of the professional forecasters. Professional forecasters' estimates of the unemployment rate were also consistent with Okun's Law given their predictions of the change in real GDP. (Pierdzioch et al., 2008).

These results suggest that we economists recognize the asymmetric behavior of the unemployment rate, but have not yet been able to develop an appropriate model that captures the asymmetries better than our judgment does.

C. Long Term Labor Market Forecasts: Methodology

The Bureau of Labor Statistics makes long-run projections of a number of variables. The variables include the size of the labor force, employment by industry and by occupation. Because the projections are for a horizon of 10 or more years, they may be evaluated differently from analyses of short-term macroeconomic predictions. For example, an evaluation of these BLS long-term projections poses three methodological issues that usually are not encountered in analyses of short-term macroeconomic forecasts.

First, no other organization made projections of these variables. Consequently, there is no benchmark for judging the BLS forecasts. Second, these projections are long-term rather than the short-term macroeconomic forecasts that have been evaluated in the past. Thus, the questions that must be addressed in such an evaluation can differ from those addressed in the macro forecasts. Finally, such a projection is a one-time forecast.[24] As an example, I will illustrate these

[24] In most forecast evaluations, there is a time series of forecasts. It is then possible to discuss the characteristics of the average forecast. This is not possible with a single observation.

issues with an example and show how the labor force, employment by industry, and occupation projections that BLS made in 1989 for the year 2000 were evaluated. (Stekler and Thomas, 2005). Although these forecasts had already been evaluated individually (Fullerton, 2003), it was possible to both ask additional questions that had not been addressed in earlier studies and to use evaluation methodologies different from those employed previously.

1. Benchmarks

There are no other forecasts that are comparable to the BLS projections, it is, therefore, necessary to *construct* a benchmark for the projections of each variable. In each case, BLS projections are compared with similar data obtained from the forecasts of these benchmarks. The benchmarks that were selected all use data that were available at the time when the BLS projections were prepared. In actuality, the benchmarks are naïve models such as: (1) projecting the latest available information; or (2) predicting that the change over the forecast period is equal to that observed over the previous time interval, which is of the same length as the forecast period.[25] The projections that are being analyzed in this article were prepared in 1988; thus the forecast period is 12 years in length. Consequently, the change from 1976 to 1988 was used as this benchmark. At a minimum, the BLS projections should be more accurate than the forecasts of these naïve models.

2. Questions in Evaluating Long-term Projections vs. Short-term forecasts.

The questions that are appropriate for evaluating the short-term forecasts have been examined in detail, but the questions that should be asked in analyzing longer run projections have not been given the same degree of attention. Because BLS projections primarily focus on long-run trends, the questions asked and the statistics used in evaluating these forecasts should be related to the primary emphasis of the forecast.

[25] These benchmarks are identical to the ones used to calculate the U coefficients in short-run forecast evaluations.

Thus, the two *basic* questions to be asked in evaluating these projections are: (1) Have the trends, specifically structural changes, been predicted correctly? (2) Were these forecasts better than those that could have been produced by a benchmark method? Additional questions concerning the sources of the errors and whether the forecasts improved overtime can also be posed.

The statistics that can answer these questions include (1) the percentage of components where the direction of change was predicted correctly; (2) dissimilarity indexes that measure the structure of the labor force, etc. (3) contingency tables that determine whether the actual and predicted directions of change are related; and (4) Spearman Rank Correlation Coefficients that measure the relationship between the predicted and actual changes of the components of an aggregate forecast.

Questions about the labor force projection are listed in Table 3.[26] These include: What is the projected size of the labor force, by age and gender? What is the growth rate of the labor force? What are the participation rates of the various groups? What is the distribution of the total labor force by age and gender? The error measures that were used in evaluating these projections are also presented in Table 3. They include the direction of error, the absolute and percentage error, the dissimilarity index, etc. The limitations of these questions and statistics are also noted.

3. Measuring structural change: Dissimilarity Indexes.

In order to determine whether the *structural changes and major trends* that occurred between 1988 and 2000 were predicted accurately, a statistic is used that directly addresses this question. The forecast of the total labor force is an aggregated estimate, and it is important to also examine the disaggregated component predictions. Such an analysis enables one to determine whether the structure of the aggregate has been predicted accurately.

[26] The questions about the employment and occupation projections are presented in the Appendix.

Table 3
Questions Asked about the Labor Force Forecasts

Questions	Accuracy Measure	Problem with Measure and/ or Question	New Question and/or Measure
What is the size of the total labor force?	Mean Absolute Error; Percentage Error: Direction of Error	Does not distinguish between census population errors and participation rate errors; standard of comparison	How much of total labor force error is the result of participation rate errors? Standard of Comparison: 1988 participation rates
What is the size of the labor force by gender etc.	Mean Absolute Error; Percentage Error, Direction of Error	Same as total labor force	Same as total labor force
What is the growth rate of the total labor force?	Error in Percentage Points	Same as total labor force	How much of the error in the growth rate forecast is the result of participation rate errors? Standard of comparison: 1988 participation rates
What are the participation rates of total labor force? Of men? Of women? By age and sex?	Error in Percentage Points, or Absolute Error/ Participation Rate; Mean Absolute Percentage error.	Does not indicate whether direction of change in participation rate was predicted, no standard of comparison.	Were the directions of change in the participation rates accurately predicted? Standard of comparison: number of changes accurately predicted vs. predictions by chance (binomial, p=0.5)
What was the distribution of the labor force by age and sex?	_____	No Standard of Comparison	Comparison Standard :Dissimilarity Index based on 1988 Distribution

If the aggregate, *X*, is predicted according to some scenario (for example, full employment), one would want to determine whether the structure is accurate even if the total is wrong. Kolb and Stekler (1992) developed a procedure for decomposing the total error into two components—where the first measures the scenario discrepancy and the second, the structural error. They calculated the proportion of the aggregate predicted and actual totals that were associated with each of the *i* components. While their analysis was based on an information content statistic, using dissimilarity indexes would yield the same result.

A dissimilarity index is a statistic that can be used to determine whether one distribution approximates another one. Specifically, it measures the amount by which the forecasted distribution would have to change to be identical to the actual distribution. The formula for the dissimilarity index is:

$$D = 0.5 \sum | (P_{fi} / P_f) - (P_{ai} / P_a) |, \text{ where} \qquad (11)$$

P_{fi} is the forecast proportion of the labor force that will be in the *i*th group, and P_f is the forecast for the total labor force. Similarly, P_{ai} and P_a are the corresponding actual data. *D* is bounded in the interval 0 to 100 percent. The smaller the value of *D*, the smaller the difference is between the predicted and actual distributions—that is, the more accurate the forecast.

The dissimilarity index for the BLS labor force projections was based on the 14 age/gender categories that had been used in 1989 to prepare the estimates for 2000. Similar dissimilarity indexes were constructed for the other distributions that serve as standards of comparison. The values of the various dissimilarity indexes are presented in the Appendix Table A3. The benchmarks were the projections based on various estimates of the population and alternative estimates of the participation rates.

The results are mixed. In some cases, the dissimilarity indexes obtained from the BLS projections are smaller (and thus more accurate) than those of the standards of comparison. In other cases, the opposite results were obtained. However, the dissimilarity index for the actual BLS forecast never exceeds 2 percent for all age/gender categories or for men and women separately. The values of the dissimilarity indexes of the standards of comparison were comparable. While there is no statistical distribution for the dissimilarity index, the BLS projection substantially predicted the structural changes that occurred in the labor force between 1988 and 2000. On the other hand, similar results were obtained from the naïve models that served as the benchmarks.

Similar procedures were used to evaluate the employment by industry and occupational projections. The BLS employment and naïve projections were again similar, but the BLS occupational estimates were more accurate than the naïve benchmark. Stekler and Thomas concluded that the accuracy of the BLS projections were comparable to the estimates obtained from naïve extrapolative methods.

4. The Applicability of the Long-Term Evaluation Methodology

The methodology that was applied in evaluating the BLS long-term projections has not been widely used. I want to show that it has a wider applicability by evaluating some lon-run Census population projections. One benefit of this analysis is the existence of multiple projections for a given date, permitting us to determine how the accuracy changes with a reduction in the forecast horizon.

The Census Bureau makes periodic forecasts of the population of the United States 5, 10 or more years into the future. These forecasts are both for the total US population and for the number of inhabitants of each of the states. There have been many evaluations of these state

forecasts. (For example, see Smith and Sinich, 1990, 1992; Campbell, 2002; Wang, 2002). In all cases, the error measures were based on the *magnitude* of the discrepancies between the projected and actual state population figures.

In addition to statistics that measure the quantitative errors, one can use the methodology that was applied to the BLS projections to these Census data. One of the purposes of a long range projection of each state's population is to provide a picture of the **distribution** of the aggregate US population among the various states. If one were only interested in knowing whether the projections captured the important trends that actually occurred, one might not be concerned with the magnitude of the errors. The accuracy of the quantitative projections of each state's total population is then not as relevant.

It is possible that the **share** of the nation's population that was in each state was predicted correctly, but that the national total and the estimates for each of the states were inaccurate by the same proportion. In that case, the projected distribution of the state populations would have exactly matched the observed distribution. Thus, the evaluation procedure that is suggested here does not focus on the specific numbers in the projections or the magnitude of the misestimates. Rather this evaluation asks whether the projected **share** of the total US population by states was similar to the actual distribution. Such an analysis enables one to determine whether the state distribution of the aggregate population was accurate even if the aggregate estimate is inaccurate.

a. Decomposing the Errors

Assume that x_t^a is the actual aggregate population of the US at time t and x_t^f is the aggregate value that was projected for time t. The error in the aggregate projection is

$$e_t = x_t^a - x_t^f \ . \qquad (12)$$

In addition it is also possible to examine the errors associated with the population projections for each of the i states. Accordingly the proportions (f_i) of the forecasted and actual (a_i) aggregated population associated with each of the i states are:

$$\boldsymbol{x}_{\boldsymbol{i,t}}^{\boldsymbol{f}} = (f_{i,t})\, \boldsymbol{x}_{\boldsymbol{t}}^{\boldsymbol{f}} \; ; \quad \boldsymbol{x}_{\boldsymbol{i,t}}^{\boldsymbol{a}} = (a_{i,t})\, \boldsymbol{x}_{\boldsymbol{t}}^{\boldsymbol{a}} \; ; \quad \sum f_{i,t} = 1, \quad \sum a_{i,t} = 1, \tag{13}$$

The forecast error for each state is

$$e_{i,t} = \; (a_{i,t})\, \boldsymbol{x}_{\boldsymbol{t}}^{\boldsymbol{a}} \; - \; (f_{i,t})\, \boldsymbol{x}_{\boldsymbol{t}}^{\boldsymbol{f}} \; . \tag{14}$$

If the aggregate forecast is absolutely accurate, the quantitative error for each state would be

$$e_{i,t} = (a_{i,t} - f_{i,t}) \; \boldsymbol{x}_{\boldsymbol{i,t}}^{\boldsymbol{a}} \, , \tag{15}$$

which is the difference between the actual and forecast proportions of the aggregate population which is in each state. The same holds true if the aggregate forecast is inaccurate. If $\boldsymbol{x}_{\boldsymbol{t}}^{\boldsymbol{a}} \neq \boldsymbol{x}_{\boldsymbol{t}}^{\boldsymbol{f}}$,

$$e_{i,t} \; = (a_{i,t} - f_{i,t}) \; \boldsymbol{x}_{\boldsymbol{t}}^{\boldsymbol{a}} + f_{i,t} (\boldsymbol{x}_{\boldsymbol{t}}^{\boldsymbol{a}} \; - \; \boldsymbol{x}_{\boldsymbol{t}}^{\boldsymbol{f}}) \tag{16}$$

Thus the quantitative forecast error for each state, $e_{i,t}$, is the sum of two components. The first represents the error in predicting the proportion of the population in each state. The second measures the error in failing to predict the aggregate correctly. In order to evaluate these long term population forecasts, we will focus on the first term, using the dissimilarity measure as our statistic.

The alternative methodology (benchmark) in this case is a naïve model, because a valid forecasting procedure should be as accurate as this type of model. In this case, we assume that the naïve projection of the states' shares of the US population for year t+h is identical to the known distribution that is available from either the Census count or from the population estimate in year t, the year from which the projection was extended.

b. Data

We evaluate the Census state population projections that were made between 1970 and 1996 for the years 1975-2005.[27] There are seven such sets of projections. The length of the forecasting horizon varied between 2 and 25 years. The naïve projections were made using the same starting points and horizons. These projections were compared either with the actual Census counts for 1980, 1990, and 2000 and or with the population estimates that the Census Bureau made for 1975, 1985, 1995 and 2005.

c. Results

The dissimilarity indexes derived from both the Census and naïve projections are presented in Table 4. The longer was the projection horizon, the larger was the size of the dissimilarity index that was associated with the projections, i.e. the less accurate the projected distribution. This result is similar to findings about the relationship between quantitative errors and the length of the horizon in short-run forecasts. As indicated above, the size of these indexes measures the amount by which the projected distribution would have to change to be identical to the actual distribution. This was less than 1% for the very short projections to more than 5% for some of the longer horizons.

Moreover, the projections seem to have improved over time. For the 5 year projections, the values of the dissimilarity indexes declined from more than 1.5% to less than 1%. The magnitude of the index for the 10 year projection made in 1970 was almost 4%; the similar numbers for the projections made in the late 1980s and 1990s were all less than 1.5%. A similar trend was observed in the more recent 20 year projections.

Nevertheless, the Census forecasts associated with the distributions of the state population forecasts are inferior to the naïve forecasts (See Table 4). In all but one case, the

[27] The data were obtained from U.S. Bureau of the Census, Current Population Reports, Series P25, Nos. 477, 735, 937, 1017, 1044, 1053, and 1111.

dissimilarity indexes associated with the naïve forecasts are smaller than the ones derived from the comparable Census projections. The exception is the five year projection made in 1975.

IV Conclusions

In providing these perspectives on forecast evaluations, I started with questions that were posed 20 years ago. While the nomenclature may have changed, we still ask the same questions today. On the other hand, the statistical and econometric foundations of our analyses have been vastly improved and new methodologies for forecasting have been developed.

Despite all these efforts, there are not many positive results to report about the quality of our forecasting techniques. We do know that combining forecasts tends to improve accuracy. We also can test for the limits of forecastability and can determine whether we have achieved that limit yet.

On the negative side, our results questioning whether the accuracy of our forecasts has improved over time are ambiguous. We still fail to predict turning points and the short run forecasts still display biases and inefficiencies. The limited amount of evidence about long-run labor market and population projections suggests that, in some dimensions, they are no better than naïve models.

However, we should not despair but rather focus on another aspect of these results. We still have immense opportunities for productive research. Let me suggest a few entries into a laundry list of possible research topics. How can we improve our forecasting models? Using these models, what are the limits of forecastability? How does one predict turning points? Why do economists prefer failing to forecast a turn that occurs rather than predicting a turn that does

happen? How are expectations (forecasts) formed? What are appropriate techniques for evaluating multivariate forecasts? for evaluating long-run predictions?

Table 4
Values of Dissimilarity Index (percentage points)
Census and Naive Forecasts

Date Projections Made	Date of Projection						
	1975	1980	1985	1990	1995	2000	2005
1970	a 1.7	a 3.9	a 5.5	a 6.5			
	b 1.7	b 3.9	b 5.5	b 6.4			
	[0.2]	[0.4]	[0.7]	[0.8]			
1975		1.8		3.7		5.3	
		[2.7]		[0.7]		[0.9]	
1980				2.5		4.2	
				[0.4]		[0.7]	
1986				1.0	1.2	1.9	2.3
				[0.2]	[0.3]	[0.4]	[0.4]
1988				a 0.6		a 2.5	
				b 0.8		b 1.7	
				[0.1]		[0.4]	
1992					0.4	1.4	2.0
					[0.1]	[0.2]	[0.4]
1996						0.8	1.2
						[0.2]	[0.3]

Notes: Numbers in [ ] are for naive (benchmark) projections.

There were two sets of projections issued in 1970 and 1988. They are denoted a and b.

APPENDIX

Table A1
Questions Asked about the Employment by Industry Forecasts

Questions	Accuracy Measure	Problem with Measure and/ or Question	New Question and/or Measure
How many people will be employed in each industry?	Percentage Error, Mean Absolute Percent Error	No standard of comparison; gives equal weight to large and small industries	Standard of comparison: Rates of Growth equal to previous rates of growth; mean weighted percent error
Which industries would have highest (lowest) employment growth rates?	Compare the number of industries projected to grow the fastest (slowest) with those that did grow fastest (slowest).	No standard of comparison; no analysis of all industries' projected and actual growth rates.	Standard of Comparison: Forecasts of fastest (slowest) growing industries from naïve model; Spearman Rank Correlation Coefficient for all industries
What is the distribution of employment by industry?	Dissimilarity Index	No standard of comparison	Standards of Comparison: Same share as in 1988 and shares based on previous growth rates.
What were the sources of the industry employment forecast errors?	Model Simulations	None	—

Source: Stekler and Thomas (2005).

Table A2

Questions About Occupational Forecasts

Question	Accuracy Measure	Problem with Question and/or Accuracy Measure	New Question and/or Measure
How many people will be employed in each occupation?	Absolute error, Absolute Percent Error	No standard of comparison; gives equal weight to large and small occupations	Standard of Comparison: Naïve model: same growth; mean weighted percent error
Which occupations will grow fastest?	Compare the number of occupations projected to grow the fastest with those that did grow fastest; distribution of growth rates by growth adjectives	No standard of comparison; no analysis of all occupations' projected and actual growth rates	Spearman rank correlation coefficient; standard of comparison not possible due to definitional changes
Which occupations will have the largest job growth?	Compare the number of occupations that were projected to have largest job growth with those that did	No standard of comparison	Standard of Comparison not possible due to definitional changes
What is the distribution of employment by occupation?	Absolute Percent Error	No standard of comparison	Dissimilarity Index: Comparison with Naïve model
What were the sources of errors?	Model Simulations	None	----------

Source: Stekler and Thomas (2005).

Table A3
Dissimilarity Indexes
Labor Force Projections

BLS Projections		**Standards of Comparison**		
		Actual Population and BLS Part. Rate	**Actual Population** and 1988 Part. Rate	**Census Population Estimate** and 1988 Part. Rate
Gender, Age	1.83	2.02	2.24	2.32
Men, Age	1.63	0.91	0.62	1.37
Women, Age	1.91	2.86	2.4	1.32

BIBLIOGRAPHY

Allen, P. G. & Morzuch, B. J. (2006). Twenty-five Years of Progress, Problems, and Conflicting Evidence in Econometric Forecasting. What about the next 25 years? *International Journal of Forecasting*, 22, 475-492.

Baghestani, H. (2008). Federal Reserve versus Private Information: Who is the best Unemployment Rate Predictor? *Journal of Policy Modeling*, 30, 101-110.

Batchelor, R. (1990), All Forecasters are Equal, *Journal of Business and Economic Statistics*, 8, 143-144.

Batchelor, R. (1997). Bias in Macroeconomic Forecasts, *International Journal of Forecasting,* 23*,* 189- 203.

Campbell, P. R., (2002), Evaluating forecast error in state population projections using Census 2000 counts, U.S. Bureau of the Census, Population Division, Working Paper Series No. 57.

Carroll, C. D. (2003). Macroeconomic Expectations of Households and Professional Forecasters, *Quarterly Journal of Economics*, 118,269-298.

Clements, M. P., Joutz, F., & Stekler, H. O. (2007). An Evaluation of the Forecasts of the Federal Reserve: A Pooled Approach, *Journal of Applied Econometrics*, 22, 121-136.

Clements, M. P. & Krolzig, H-M. (2003), Business Cycle Asymmetries: Characterization and Testing Based on Markov-Switching Autoregressions, *Journal of Business and Economic Statistics*, 21, 196-211.

Davies, A. & Lahiri, K. (1998). Examining the Rational Expectations Hypothesis Using Data on Multiperiod Forecasts. In Analysis of Panels and Limited Information Dependent Models, ed. C. Hsiao, et al. Cambridge University Press, 226-254,

Diebold, F.X. & Mariano, R. (1995). Comparing Predictive Accuracy, *Journal of Business and Economic Statistics*, 13, 253-265.

Diebold, F. X. & Rudebusch, G. D. (1989). Scoring the Leading Indicators, *Journal of Business*, 62, 369-391.

Diebold, F. X. & Rudebusch, G. D. (1991). Forecasting Output with the Composite Leading Index: A Real Time Analysis, *Journal of the American Statistical Association*, 86, 603-610.

Dopke, J. & Fritsche U., (2006) , Growth and inflation forecasts for Germany, *Empirical Economics*, 31, 777–798.

Dougherty, J. (2009). The U6 U3 Difference as a Turning Point Indicator, George Washington University, mimeo

Fair, R. C. , (2009) Analyzing Macroeconomic Forecastability, (mimeo)

Fildes, R. & Stekler, H. O. (2002). The State of Macroeconomic Forecasting, *Journal of Macroeconomics*, 24, 435-468.

Fullerton, H. N. (2003). Evaluating the BLS Labor Force Projections to 2000, *Monthly Labor Review*, October, 3-12.

Golan, A. & Perloff, J. M. (2004). Superior Forecasts of the U.S. Unemployment Rate Using a Nonparametric Model, *Review of Economics and Statistics*, 86, 433-438.

Elliott, G. & Timmermann, A. (2008). Economic Forecasting, *Journal of Economic Literature*, 46, 3-56.

Harvey, D., S. Leybourne and P. Newbold (1997). "Testing the Equality of Prediction Mean Squared Errors," *International Journal of Forecasting*, 13, 281-91.

Heilemann, U. and Stekler, H.O. (2003), Has the accuracy of German macroeconomic forecasts improved? Discussion paper of the German Research Council's research centre, 475, 31/03.

Heilemann, U. (2002). Increasing the Transparency of Macroeconometric Forecasts: A Report from the trenches, *International Journal of Forecasting*, 18, 85-105.
Henriksson R.D. & Merton R.C. (1981): On the Market Timing and Investment Performance of Managed Portfolios II - Statistical Procedures for Evaluating Forecasting Skills, Journal of Business, 54, 513-533.

Hendry-Richard

Holden, K. & Peel D.A. (1990). On Testing for Unbiasedness and Efficiency of Economic Forecasts, *Manchester School*, 58, 120-127.

Isiklar, G. & Lahiri, K. (2007). How far ahead can we Forecast? Evidence from Cross-Country Surveys, *International Journal of Forecasting*, 23, 167-187.

Joutz, F. & Stekler, H.O. (2008) Another look at the Fed Forecasts, George Washington University, mimeo.

Kaylen, M.S. & Brandt, J. A., (1988), A note on qualitative forecast evaluation: Comment, *American Journal of Agricultural Economics,* 70, 415-16.

Kolb, R. A. & Stekler, H. O., (1992), Information Content of Long-term Employment Forecasts, *Applied Economics*, 24, 593-596.

Lawrence, M., Goodwin, P., O'Connor, M. & Onkal, D. (2006). Judgmental Forecasting: A Review of Progress over the last 25 Years, *International Journal of Forecasting*, 22, 493- 518.

Leitch, G & Tanner, E.J., (1991). Economic forecast evaluations: Profits versus conventional error measures, *American Economic Review,* 81, 580-590.

Marcellino, M. (2006). Instability and Non-Linearity in the EMU. In C. Milas, P. Rothman, & D. van Dyck (Eds) *Nonlinear Time Series Analysis of Business Cycles*, (pp.151-174). Amsterdam: Elsevier.

Merton R.C. (1981). On Market Timing and Investment Performance of Managed Performance I - An Equilibrium Theory of Value for Market Forecasts, *Journal of Business*, 5, 363-406.

Milas, C. & Rothman, P. (2008). Out-of-Sample Forecasting of Unemployment Rates with Pooled STVECM Forecasts, *International Journal of Forecasting*, 24, 101-121.

Mincer, J. & Zarnowitz, V. (1969). The Evaluation of Economic Forecasts. In *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, ed. J. Mincer, NBER, 14-20.

Montgomery, A. L., Zarnowitz, V., Tsay, R. S., & Tiao, G. C. (1998). Forecasting the U.S. Unemployment Rate, *Journal of the American Statistical Association*, 93, 478-493.

Moshiri, S. & Brown, L. (2004). Unemployment Variation over the Business Cycles: A Comparison of Forecasting Models, *Journal of Forecasting*, 23, 497-511.

Naik, G. & Leuthold, R. M., (1986), A note on qualitative forecast evaluation, *American Journal of Agricultural Economics,* 68, 721-26.

Neftci, S. N. (1984). Are Economic Time Series Asymmetric Over the Business Cycle? *Journal of Political Economy*, 92, 307-328.

Newey, W. and K. West (1987). "A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

Oller, L-E. & Barot B. (2000). Comparing the Accuracy of European GDP Forecasts, *International Journal of Forecasting,* 16, 293-315.

Oller, L-E. & Teterukovsky, A. (2007). Quantifying the Quality of Macroeconomic Variables, *International Journal of Forecasting*, 23, 205-217.

Orphanides, A. (2001). Monetary Policy Rules Based on Real-Time Data, *The American Economic Review*. 91(4), 964-985.

Pesaran, M. H. & Skouras, S. (2002). Decision-Based methods for Forecast Evaluation, In *A Companion to Economic Forecasting*, ed. M. Clements and D. Hendry, Blackwell,24-267.

Pesaran, M. H. & Timmermann, A. (2004). How Costly is it to Ignore breaks when Forecasting the Direction of a Time series? *International Journal of Forecasting,* 20, 411-425.

Pierdzioch, C., Rulke, J-C, 7 Stadtmann, G. (2008). Do Professional Economists' Forecasts reflect Okun's Law? Some Evidence for the G-7 Countries. (mimeo).
Pool, D. A. and Speight, A. E. H. (2000). Threshold Nonlinearities in Unemployment Rates: Further Evidence for the UK and G3 Economies, *Applied Economics*, 32, 707-715.

Romer, C. D. and Romer D. H.. (2008). The FOMC versus the Staff: Where Can Monetary Policymakers Add Value? *American Economic Review* 98(2), pp. 230-35.

Rothman, P. (1998). Forecasting Asymmetric Unemployment Rates, *Review of Economics and Statistics*, 80, 164-168.

Schnader, M. H. & Stekler, H. O. (1990). Evaluating Predictions of Change, *Journal of Business*, 63, 99-107.

Seip, K.L. & McNown (2007). The Timing and Accuracy of Leading and Lagging Business Cycle Indicators, *International Journal of Forecasting*, 23, 277-287.

Sinclair, T. M., Stekler, H.O. & Kitzinger, L. Directional Forecasts of GDP and Inflation: A Joint Evaluation with an Application to Federal Reserve Predictions, *Applied Economics* (forthcoming).

Sinclair, T. M., Gamber, E. N., Stekler, H. O. & Reid, E., Jointly Evaluating GDP and Inflation Forecasts in the Context of the Taylor Rule, (mimeo).

Skalin, J. & Terasvirta, (2002). Modelling Asymmetries and Moving Equilibria in Unemployment Rates, *Macroeconomic Dynamics*, 6, 202-241.

Smith, S. K. & Sinich, T., (1990), The Relationship between the Length of the Base Period and Population Forecast Errors, *Journal of the American Statistical Association*, 85, 367-375.

Smith, S. K.& Sinich, T., (1992), Evaluating the Forecast Accuracy and Bias of Alternative Population Projections for States, *International Journal of Forecasting*, 8, 495-508.

Stekler, H.O. (1989). Turning Point Predictions, Errors and Procedures", in Kajal Lahiri and Geoffrey H. Moore, eds. *Leading Economic Indicators: New Approaches and Forecasting Record,* Cambridge University Press.

Stekler, H. O. (1991). Macroeconomic Forecast Evaluation Techniques, *International Journal of Forecasting*, 7, 375-384.

Stekler, H. O. (1994). Are Economic Forecasts Valuable? *Journal of Forecasting*, 13, 493-505.

Stekler, H. O. (2001). The Rationality and Efficiency of Individuals' Forecasts, In *A Companion to Economic Forecasting*, ed. M. Clements and D. Hendry, Blackwell, 222-240.

Stekler, H.O. (2008). What do we Know about G-7 Macro Forecasts?.....

Stekler H.O. & Thomas, R. (2005). Evaluating BLS Labor Force, Employment, and Occupation Projections for 2000, *Monthly Labor Review*, July, 46-56.

Swanson, N. R. & White H. (1997). A Model Selection Approach to Real-Time Macroeconomic Forecasting Using Linear Models and Artificial Neural Networks, *Review of Economics and Statistics*, 79, 540-550.

Taylor, J.B. (1993). Discretion versus Policy Rules in Practice, *Carnegie-Rochester Conference Series on Public Policy,* 39,195-214.

Timmerman, A., 2007, An evaluation of the World Economic Outlook, *IMF Staff Papers*, 54, 1–33.

Vogel, L., 2007. How do the OECD growth projections for the G7 economies perform? OECD Economics Department Working Papers No. 573.

Vuchelen, J. & Hutierrez, M-I, (2005). Do the OECD 24 Month Horizon Growth Forecasts for the G-7 Countries Contain Information? Applied Economics, 37, 855-862.

Wang, C., (2002), Evaluation of Census Bureau's 1995-2005 state population projections, Working Paper no. 67, US Census Bureau.

West, K. D. (2006). Forecast Evaluation, in Handbook of Economic Forecasting Vol. 1, ed. G. Elliott, C. W. J. Granger & A. Timmermann, North Holland, 99-134.

Woodford, M. (2001a). The Taylor Rule and Optimal Monetary Policy, *American Economic Review*, 91 (2), 232-237.

Woodford, M. (2001b). Inflation Stabilization and Welfare, National Bureau of Economic Research working paper 8071.