Journal of Macroeconomics 24 (2002) 435–468

Journal of
MACROECONOMICS

www.elsevier.com/locate/econbase

The state of macroeconomic forecasting

Robert Fildes [a], Herman Stekler [b,*]

[a] *Department of Management Science, Lancaster University, Lancaster, LA1 4YX, UK*
[b] *Department of Economics, George Washington University, 2201 G. Street, NW, Washington, DC 20052, USA*

Received 21 December 1999; accepted 6 March 2002

Abstract

Macroeconomic forecasts are used extensively in industry and government. The historical accuracy of US and UK forecasts are examined in the light of different approaches to evaluating macroforecasts. Issues discussed include the comparative accuracy of macroeconometric models compared to their time series alternatives, whether the forecasting record has improved over time, the rationality of macroeconomic forecasts and how a forecasting service should be chosen. The role of judgement in producing the forecasts is also considered where the evidence unequivocally favors such interventions. Finally the use of macroeconomic forecasts and their effectiveness is discussed. The conclusion drawn is that researchers have paid too little attention to the issue of improving the forecasting accuracy record. Finally, areas where improvements are most likely to be found are discussed.

JEL classification: E37
Keywords: Econometric models; Macroeconomic forecasting; Rationality; Forecast evaluation; Structural breaks; Cliometrics; Industry structure

Considerable intellectual activity is devoted to forecasting major economic variables. It is therefore important to determine how well we have performed this activity and what lessons may lead to improvements. Given the nature of the macroeconomic forecasting industry (as surveyed by Fildes, 1995), it is impossible to review all aspects of the field. We will, therefore, only examine the predictions of short-run real GDP and inflation forecasts for the US and the UK, citing other countries' data only to confirm (or contradict) our findings. We focus on these variables because they are

[*] Corresponding author. Tel.: +1-202-994-6150; fax: +1-202-994-6147.
E-mail address: hstekler@gwu.edu (H. Stekler).

0164-0704/02/$ - see front matter © 2002 Elsevier Science Inc. All rights reserved.
PII: S0164-0704(02)00055-1

of interest to the entire profession. Unfortunately, this leaves many questions unanswered, e.g. which GDP components were the hardest to predict and which contributed most to the inaccuracy of the aggregate forecasts?

We survey the forecasts generated by a variety of techniques and the contribution that expert judgment has made to the modeling process. Since others have compared the structure of models used in macroeconomic forecasting (see for example, Klein, 1991; Wallis, 1993) and dealt with the history and structure of the industry (Bodkin et al., 1991; Daub, 1987; Smith, 1994; Den Butter and Morgan, 1998), we only deal with the forecasting record of these models.

An economic forecast should provide information about (1) the economy's direction of movement, (2) the timing of any turning point, (3) the magnitude of the change and (4) the length of time over which a movement is expected to persist. Forecast evaluations based on these dimensions have asked a number of questions. What is the magnitude of the average forecast error (relative to the outcome)? Were the dates of the recessions predicted accurately? Were the major inflationary cycles forecast? Was there a bias in the predicted direction of change? Why did the errors occur? Answers to these questions provide the users of forecasts with an understanding of the strengths and limitations of the predictions. The forecasters require these answers in order to improve the quality of their output.

In addition, we address other topics: (1) Do judgmental adjustments improve the accuracy of the econometric and time series predictions? (2) Is one method or forecaster consistently superior to another? (3) Are the forecasts superior to readily available and simpler benchmarks? (4) Is it desirable to combine forecasts from different sources?

This survey will report on the findings of previous evaluations of macroeconomic forecasts and the factors that have affected accuracy. A secondary aim is to develop an agenda for improving macroeconomic forecasting. The first section of the paper deals with problems associated with measuring accuracy—the data, descriptive statistics, standards of comparison and rationality tests. Section 2 examines the forecasting record, while the next section judges the forecasts. We then question whether the forecasts were optimal and examine two techniques for improving forecast accuracy: combining predictions and making judgmental adjustments. The final section analyzes new methodologies for improving predictive accuracy and presents suggestions for further research.

1. Measuring accuracy

1.1. The base line data

Because GDP data are revised periodically, it matters which vintage of the outcome data is used in any forecast evaluation. Using McNees' data (1976a, 1979, 1985) we can show how different data 'vintages' affect the size of forecast errors. For example, in 1971.2–1976.4, the mean absolute error (MAE) of the real GNP predictions varies between 1.55% and 1.86%, depending on which data are used. For the

GNP deflator, the similar range is 1.70–2.02%. [1] The differences occur because McNees uses the data that are available at the time that the evaluation is undertaken as his measure of the outcomes. If the revised data show that GNP grew at x% rather than the preliminary published estimate of y%, the forecast error would be changed by the difference $x - y$.

One can argue that the results of forecast evaluations should not change with each revision of the National Income Accounts. Zarnowitz, therefore, always uses one vintage of actual data with which the forecasts are compared. These are the first revised figures, published 45 days after the end of the quarter to which they refer. Zarnowitz argues that individuals should not be judged on their ability to predict how the statistical agencies will revise the data in the future. In contrast to the US studies, most authors who evaluated the forecasts of other countries generally did not indicate which vintage of data sets were used in their analyses. Melliss' (1997) and Cho's (2002) reviews of the UK and US forecasts respectively found no consistent effect from the vintage of the data.

The purpose of the forecast should determine how it should be judged and thus which set of actuals should be used. Analysts concerned with financial markets would like the forecasts to accurately predict the preliminary figures because the publication of these numbers might affect the financial markets. On the other hand, decision makers would like to know what "actually" will happen (as measured by the accuracy of the final numbers).

1.2. Qualitative measures

The choice of measures for evaluating macroeconomic forecasts depends upon the questions that are under investigation. The qualitative failures of forecasts include turning point and systematic errors. Evaluations would determine the number of cyclical turns that were predicted, the number that were not forecast, and the number of false turns that were predicted. The search for systematic errors would consider both the number and magnitude of over and underestimates of the variable and determine whether the errors are associated with particular phases of the business cycle.

1.3. Descriptive statistics

To evaluate quantitative forecasts, quantitative measures that summarize the raw errors must be used. Let $\widehat{Y}_t(k)$ represent the k period ahead forecast of the variable of interest, Y_{t+k}, made at time t for period $t + k$, and $e_Y(t, k)$ is the corresponding k period ahead forecast error. It will then be possible to compare different methods and forecasters and to determine whether there are any statistically significant differences among them. Among the most frequently used error measures are: (1) the mean

[1] These date are available from the authors.

error, ME, that measures bias. For lead time k based on time origins $T_0 + 1, \ldots, T_0 + n$ it is defined as

$$\mathrm{ME} := \frac{1}{n}\sum_{1}^{n} e_Y(T_0 + j, k);$$

(2) the root mean square error (RMSE),

$$\mathrm{RMSE} := \sqrt{\frac{1}{n}\sum_{j=1}^{n} e_Y^2(T_0 + j, k)};$$

and (3) MAE,

$$\mathrm{MAE} := \frac{1}{n}\sum_{j=1}^{n} |e_Y(T_0 + j, k)|.$$

All these measures have a number of limitations.

First, by themselves, they do not provide meaningful information and, moreover, they have been subject to intense criticisms. The RMSE is particularly affected by outliers that are common in economic data. Neither of the measures is naturally scale independent except when applied to percentage changes (Armstrong and Collopy, 1992; Fildes, 1992) and they involve averaging errors over observations that have different degrees of variability (Fair, 1980; Jenkins, 1982; Diebold, in Baillie et al., 1993). The issue of scaling is also critical when a number of data series are being analyzed. [2] The aim in adopting particular error measures is to provide a simple and fair characterization of the distribution of the errors.

Despite such criticisms, the RMSE and MAE have invariably been used as standards for judging the quality of predictions. For example, the RMSE is compared with the standard deviation of the series being forecast and the MAE is compared with the mean absolute change of the predicted variable. These measures show how the errors are related to the variability of the series that is being forecast.

Typically, these error measures have focussed on the point forecast alone. Recently, attention has also been given to the uncertainty around the published forecasts. The question of whether estimates of the uncertainty in a point forecast are well calibrated or, more generally, the estimated probability distribution matches the realized distribution (Tay and Wallis, 2000) has received relatively little attention as there are little data available but see for example, Diebold et al. (1999). [3]

[2] Clements and Hendry (1993) have proposed a generalization of the RMSE that takes into account the intercorrelations between errors when more than one macroeconomic series is being analyzed and to ensure invariance to linear transformation. The practicality of their measure has been questioned because typical macroeconomic forecast evaluations are based on small sample of non-normal data (Armstrong and Fildes, 1995). There has been some discussion whether invariance is necessarily a desirable measure (Baillie et al., 1993). Although there are several articles that deal with the distribution of errors (e.g. Diebold et al., 1998), this issue has been disregarded in most forecast evaluations.

[3] For many users an accurate measure of future uncertainty may be as valuable as the point forecast itself and this poses a second order forecasting problem (Tay and Wallis, 2000). In general, Chatfield (1993) has argued, both model based and judgmental estimates tend to underestimate the uncertainty in forecasts.

1.4. Naïve and time series benchmarks

A better approach for evaluating performance is to compare these descriptive statistics with similar statistics obtained from a naïve standard or simple time series model. Using a standard permits one to test whether a forecaster's errors are significantly smaller than those of the benchmark. A frequently used naïve standard compares a forecaster's errors with those obtained from a no-change (random walk) naïve model. Theil's (1966) U coefficient formalizes this comparison.

$$U = \sqrt{\frac{\sum (Y_{t+k} - \widehat{Y}_t(k))^2}{\sum Y_{t+k}^2}}.$$

If U is less than 1, the forecasts which are being evaluated have smaller errors than those of the naïve model, *but this result does not guarantee that the former are significantly better than the latter.* This naïve model (N1) serves as an appropriate minimum standard of comparison for variables that experience both positive and negative changes. However, a different standard is required for variables such as the price level, which almost always show positive rates of growth. In this case, the naïve model (N2) extrapolates the same change as was observed in the last period, and the errors of N2 are used in the denominator of U.

The naïve models establish the minimum level of accuracy that a set of forecasts should have. Moore (1969) suggested that the accuracy of forecasts should be compared with the predictions obtained from time series extrapolations. From the 1970s on evaluations have adopted even more demanding benchmarks by using ARIMA or VAR time series models as appropriate standards (Holden, 1995). The rationale is that forecasters should perform at least as well as the simplest time series models from which predictions could have been derived (Nelson, 1972, 1984). There is, however, a problem with using these benchmarks. They are usually estimated with the latest available data, and not the numbers that existed when the forecasts were made.

Whatever benchmark is used in the evaluation of forecasts, the difference between the two sets of errors should be tested for statistical significance. Granger and Newbold (1977) had proposed a useful statistic for making MSE comparisons. A new statistic developed by Diebold and Mariano (1995) permits a statistical comparison when other quantitative error measures are reported. This test was further extended by Harvey et al. (1997) and Ashley (1998) has developed a new measure, based on bootstrap inference, that does not require large samples.

1.5. Rationality tests

A comparison of forecasts with benchmark standards determines which one has smaller errors. It does not indicate how to improve upon the observed record. Rationality tests determine whether or not the predictions are optimal with regard to a particular information set (Wallis, 1989). A variety of rationality tests have been proposed. The simplest tests for bias and weak form informational efficiency and is based on the regression

$$Y_{t+k} = \alpha + \beta \widehat{Y}_t(k) + e_{t+k}$$

where Y_t and $\widehat{Y}_t(k)$ are the actual and k period ahead predicted values at time t respectively. The joint hypothesis is that the coefficients α and β do not differ significantly from 0 and 1, respectively. A second test is that the errors are MA $(k-1)$. For $k=1$, the errors must not be serially correlated. A rejection of these hypotheses would indicate that the forecasts are biased or have not incorporated the information contained in past errors. [4] Since the joint hypothesis (of $\alpha = 0$, $\beta = 1$) is sufficient for unbiasedness but not a necessary condition (Holden and Peel, 1990), a more restrictive condition is to test

$$Y_{t+k} - \widehat{Y}_t(k) = \alpha + e_{t+k}$$

with null hypothesis: $\alpha = 0$.

An improvement in accuracy is also possible if the forecast error is correlated with *any* information that was known at the time that the forecast was prepared. A correlation indicates that past information has not been used efficiently. This test is based on the regression

$$e_Y(t-k,k) = \gamma + \sum_i \sum_{j>k} \delta_{i,j} X_i(t-j) + e_t$$

where the Xs are variables, which might have affected the forecast. The joint hypothesis is that γ and all the $\delta_{i,j}$ are not significantly different from zero. A rejection of this hypothesis indicates the possibility of improving forecast accuracy by using the information contained in those variables, which had non-zero coefficients. While the rejection of these rationality tests suggests that forecasts might have been improved, a caveat should be noted. Jeong and Maddala (1991) show that tests of rationality depend upon the statistical assumptions and procedures that are used. For example, the regression tests are not valid in the presence of unit roots, and cointegration tests should be used (Pain, 1994). In addition, it is also possible that the forecaster's loss function is asymmetric and the predictions would still be rational even if the ME is non-zero.

2. The record

In this section we summarize the results of the many studies that have evaluated the accuracy of macroeconomic forecasts. We examine both the turning point errors and various quantitative measures of forecast accuracy.

[4] For fixed outcomes the test for weak form efficiency is that forecast revisions are independent of past available information (Nordhaus, 1987; Clements, 1997).

2.1. Turning point forecasts; systematic errors

2.1.1. The US forecasts

Zarnowitz (1967, 1979, 1984, 1992), Zarnowitz and Braun (1992), McNees (1976b, 1986, 1988, 1992a) and McNees and Ries (1983) have provided the most complete analyses of US macroeconomic forecasts. Zarnowitz analyzes the record of professional macroeconomic forecasters responding in the ASA/NBER surveys, latterly called the Survey of Professional Forecasters. McNees examines the predictions of various econometric models, but his analyses also include the median ASA/NBER predictions. The results are robust with respect to both the covered time periods and the predictive methods.

A disturbing finding is that recessions were generally not forecast in advance. This is true for both annual and quarterly predictions (McNees, 1976b, 1988, 1992a; McNees and Ries, 1983; Zarnowitz, 1992). For example, both the 1974 and 1981 peaks were not recognized even as they occurred. While missing those points, the forecasts did predict slowdowns in the economy. The forecasters were thus successful in distinguishing between periods of rapid growth on one hand and slow and negative growth on the other hand. Forecasters only recognized that the 'slowdowns' had turned into recessions after the quarter in which the economy had its peak (Schnader and Stekler, 1990; McNees, 1992b; Stekler, 1994).

Economists made few predictions of peaks that did not occur, with the false turns of 1978–79 an exception. [5] On the other hand, they were willing to forecast the end of recessions (Zarnowitz, 1992), even to the extent of predicting such turns too quickly (McNees, 1976b, 1987). The factors that produce such asymmetries have not been adequately explored. [6]

Forecasters have also made systematic errors. They overestimated the rate of growth during slowdowns and recessions and underestimated it during recoveries and booms (Zarnowitz and Braun, 1992). It had been argued that the underestimates of growth resulted from optimal forecasts (Mincer and Zarnowitz, 1969; Hatanaka, 1975; Samuelson, 1976) because the variance of the predictions *should* be less than that of the outcomes. However, the empirical evidence has not been consistent with this assumption (Stekler, 1975; Smyth and Ash, 1981). One explanation for these underestimates is that forecasters failed to recognize the strength of booms because they used a 'regression towards the mean' heuristic (Zarnowitz, 1979).

[5] While there was no recession in 1978–79, real GNP declined for two quarters during this period. Thus, the economists foresaw economic weaknesses but recessions did not develop. The question that may be asked is why forecasters were so willing to predict a recession in this period. Economists had badly missed the date of the previous (1973–75) recession that was associated with an oil shock and did not wish to make a similar error when the second oil shock occurred. The forecasters over-compensated for this last observed error, an example of the so-called 'anchoring heuristic' often observed when human judgment is used to supplement model-based forecasts (Bolger and Harvey, 1998).

[6] Stekler (1972) suggested that forecasters had zero priors that a recession would occur. Differential costs may also have been assigned to the two types of turning point errors: predicting a recession that did not occur and failing to predict one (Schnader and Stekler, 1998).

The forecasts of the US inflation rate also had systematic errors. Inflation was generally underpredicted when it was rising and overpredicted when it was declining (Zarnowitz and Braun, 1992; McNees and Ries, 1983). Particularly large errors occurred during the periods when prices were rising rapidly during the 1970s and early 1980s. Diebold et al. (1998) provided additional evidence about the systematic errors in the predictions in the Survey of Professional Forecasters. These forecasters provide a probability density function along with point forecasts of inflation and their average estimates were not well calibrated.

These qualitative results indicate that forecasters made systematic errors in predicting both the rate of growth and the inflation rate of the US economy. These errors occurred when the economy was subject to major perturbations, just the times when accurate forecasts were most needed. We feel there are few excuses to be made for the persistent, systematic errors we observe. Forecasters should be expected to forecast recessions (but see Kenneth West's commentary). Leading indicators, while predicting too many downturns, generally anticipate these recessions. This information is available to the forecasters. On the other hand, the economists cannot be expected to forecast OPEC oil shocks, although predicting the effects of such shocks are their responsibility.

2.1.2. UK forecasts and other evidence

There have been fewer evaluations of UK forecasts. Forecasts are issued by the government, by a number of research organizations, and by a substantial number of commercial forecasting organizations. The UK studies did not specifically determine whether the forecasters had predicted cyclical turning points in advance; rather the analyses have focussed on the magnitude of the forecast errors. The limited evidence suggests that forecasters were not able to forecast turns (Barker, 1985; Wallis, 1989; Britton and Pain, 1992). Mills and Pepper (1999) confirm this result and find that the UK forecasters not only failed to predict the cyclical turns of the 1979–82 and 1989–92 recessions in advance, but, in addition, were only partially successful in identifying a turn when it occurred.

The inflation surges of 1979–80 and 1989–90 were not predicted until prices had been rising rapidly for a number of months; a similar result was observed when inflation tapered off (Mills and Pepper, 1999). As in the US there is a general tendency to overpredict inflation when it is falling and to underpredict it when it is rising (Wallis, 1989; Melliss, 1997).

As for other countries, Ash et al. (1998) analyzed whether the OECD forecasts, made with horizons of 6, 12, and 18 months, correctly predicted the direction of change of GNP (see also Pons, 2000). They showed that few of the 12–18 month ahead forecasts were significantly better, in terms of predicting the direction of change, than the naïve no change model. This conclusion was confirmed by Öller and Barot (2000). Further detailed evidence is provided by Blix et al. (2001). In addition Anderson (1997) indicated that the OECD predictions tended to underestimate changes in the trends of both real output growth and inflation. However, in periods when inflation was decelerating, the rate of inflation was overpredicted.

We can thus conclude that the UK and OECD evidence are in accord with the findings that were obtained from the evaluations of the US forecasts.

2.2. Quantitative results

Although the qualitative findings about turning points and systematic errors are crucial to understanding the forecasting record over cyclical periods, most evaluations have provided quantitative measures.

2.2.1. US forecasts

The US forecasts not only cover different time periods, but also involve estimates over different horizons, i.e. annual and quarterly spans from 0 to *n* periods (e.g. the growth over the next 4 quarters). In the latter category, both the forecasts made over a span of 0–*n* quarters and the predictions made for a particular quarter *k* periods in advance have been examined (e.g. the growth in quarter 4). [7]

2.2.1.1. One-year forecasts. Annual forecasts of real GNP growth, made for various sub-periods of 1962–89, had MAEs slightly in excess of 1% of GNP. These errors were at least 25% of the mean absolute change in GNP (Zarnowitz, 1992). The Michigan model's forecasts had similar errors (McNees, 1988). Similarly, Zarnowitz showed that the inflation predictions had MAEs of 1.0–1.4%, which, depending on the time period, should be compared with mean absolute changes of 4.2–5.9%.

2.2.1.2. Quarterly estimates and forecasts. There are several questions involving the quality of multi-period quarterly forecasts: What is the magnitude of the errors? What is the relationship between the length of the forecasting lead and the size of the errors? The results show that the accuracy of the predictions of GNP growth improve as the length of the prediction horizon decreases. *A substantial improvement occurs when the task of forecasting switches from predicting what will happen in the next quarter to estimating the level of activity of the current period.* Table 1 shows that the MAE of the median ASA/NBER forecast of the change in real GNP for the current quarter was 2.36% (annual rate calculated according to Footnote 7) rising to 3.04% (intra-quarter change 1–2) for the subsequent period and 3.68% for the change expected three-quarters in advance. Similar results hold for the inflation predictions. When predictions are made every month, accuracy improves substantially as the actual data for the first month of the current quarter becomes available (McNees, 1988; Kolb and Stekler, 1990).

These results are not surprising. The accuracy of forecasts made three quarters ahead is not likely to be significantly better than the prediction made a year in advance unless there is substantial new information. That is unlikely. The availability of actual data for the current period provides a significant improvement in accuracy, but these projections are not pure forecasts. Rather they are a combination of

[7] A difficulty arises in interpreting the results of the major published studies. Although both McNees and Zarnowitz transform their data into growth rates, McNees converts these data into annual rates of growth; Zarnowitz does not make this conversion. An approximate adjustment to the published figures in the latter's studies is to multiply all quarterly errors by 4; all semiannual data by 2; etc.

Table 1
MAE of median ASA/NBER forecasts of GNP, real GNP and the GNP price deflator, span and intra-quarter forecasts (% change, annualized rates in parentheses)

Variable	MAE								
	Span					Intra-quarter changes			
	0–1	0–2	0–3	0–4	0–5	1–2	2–3	3–4	4–5
GNP	0.56	1.02	1.49	1.84	2.31	0.73	0.81	0.82	0.86
	(2.24)	(2.04)	(1.99)	(1.84)	(1.85)	(2.92)	(3.24)	(3.28)	(3.44)
Real GNP	0.59	1.00	1.41	1.82	2.26	0.76	0.82	0.92	0.94
	(2.36)	(2.00)	(1.88)	(1.82)	(1.81)	(3.04)	(3.28)	(3.68)	(3.76)
Price deflator	0.38	0.72	1.08	1.55	2.07	0.49	0.54	0.57	0.62
	(1.52)	(1.44)	(1.44)	(1.55)	(1.66)	(1.96)	(2.16)	(2.28)	(2.48)

Source: Zarnowitz and Braun (1992).

estimates based on the newly available data and forecasts of the activities for which information is not yet available. [8]

Table 1 indicates that the increase in accuracy as the forecasting horizon decreases is not monotonic. This is especially true for the forecasts made over a span of *n* quarters. The longer span forecasts of real GNP were more accurate than the shorter ones. One reason that longer span forecasts might be more accurate than the shorter span ones is that the former capture the trends that prevail over the longer period but do not precisely determine the dynamics of the quarterly movements.

2.2.2. *The UK and other evidence*

Most UK forecast evaluations include the UK Treasury predictions (Burns, 1986; Treasury and Civil Service Committee, 1991; Britton and Pain, 1992; Melliss, 1997; Melliss and Whittaker, 1998). The 1991 study compared the accuracy of the Treasury forecasts with those available from other major providers. [9]

Based on year ahead forecasts, Table 2 shows that the MAE of the Treasury's forecasts of real GDP growth was 0.8% and 1.00% in 1986–90 and 1990–98, respectively. The MAE was about 60% of the mean absolute change (the Naïve forecasting) in both periods. The non-Treasury errors were slightly larger in the first period but smaller in the second one. The inflation errors were similar for the latter period of relatively low inflation but the increasing inflation, 1988–90 was very poorly forecast. These results are similar to the findings obtained from a comparison of the Treasury

[8] In order to improve the accuracy of quarterly forecast, model-based approaches for incorporating current monthly information have been proposed. (Howrey, 1978; Corrado and Greene, 1988).

[9] The different forecasting groups have been categorized into 'Independent'—typically units with an academic, not-for-profit institutional base, 'City'—those forecasting groups based in the City of London, and 'Selected independents'—a treasury selected sub-set chosen to reflect established opinion and includes some City-based forecasters. The two consensus forecasts are calculated from the average forecasts of the respective groups.

Table 2
One year ahead UK forecast errors—mean (median) absolute error

Forecasting group	GDP		Inflation	
	1986–90	1990–98	1986–90	1990–98
Independent average	1.2	0.95 (0.91)	1.6	1.21 (0.80)
Selected independents	1.0	0.87 (0.72)	1.4	1.23 (0.71)
Independent consensus	NA	0.89 (0.96)	NA	1.11 (0.70)
City average	1.0	0.85 (0.75)	1.7	1.23 (0.75)
City consensus	NA	0.82 (0.96)	NA	1.06 (0.60)
Treasury	0.8	1.00 (0.79)	1.5	1.01 (0.55)
Average outcomes	3.05	1.59	6.32	3.72
Standard deviation outcomes	1.61	1.96	2.68	2.47
Naïve forecast	1.35	1.60	1.74	1.70

Source: UK Treasury compilation of forecasts for the 1990–98 calculations and Treasury and Civil Service Committee, 1991), forecasts made for the March budget.

NB: GDP is based on preliminary figures, average estimates of GDP. RPI (retail price index, 4th quarter) is used for inflation apart from those forecasting groups who forecast averaged inflation over the year. Since 1993 the Treasury have only forecast RPI (excluding interest charges on housing—MIPS) and this has been used to calculate the treasury error statistics. Last year's actuals have been used in the calculation of the Naïve forecast.

forecasts with those of the National Institute, the CBI (Confederation of British Industry) and London Business School forecasts (Britton and Pain, 1992).

Systematic errors are made in forecasting inflation. It is overpredicted when low and underpredicted when high. Overwhelmingly, most forecasters make the same mistakes. For example, around 90% of forecasters overpredicted growth in 1990–92. Similarly, 90% overpredicted inflation for 1991–94. No single year since 1990 saw the forecasters getting both variables broadly right. There is little or no correlation between performance on GDP and performance on inflation, a finding in accord with McNees' (1992a). Moreover, neither of the consensus forecasts did particularly well. *We conclude that despite adhering to different economic theories and using different methodologies these forecasters make the same mistakes—or at the very least, miss out on the same large shocks, a result in accord with US evidence.*

For thirteen European countries, Öller and Barot (2000) examined the GDP record of the national forecasting institutes in Europe as well as the OECD forecasts made for the same countries. The accuracy of the two sets of predictions was similar. For the period 1971–95 the average absolute error is 1.43% and ranges from 1.05% for France to 2.12% for Finland.

2.2.2.1. Quarterly estimates. The major information about quarterly UK forecasts comes from the Treasury predictions (Table 3).

Using the MAE as a measure, the Treasury forecasts exhibit a pattern similar to the one that was observed in the US predictions. While the absolute forecast error increases with the forecast span for both GDP and the RPI, the annualized errors decrease for GDP while increasing for RPI. These results are compatible with the evidence of Holden and Peel (1985) who analyzed the National Institute Forecasts.

Table 3
Quarterly error statistics for the UK Treasury (annualized rates): 1971–96

Span of forecast horizon	GDP		Inflation (RPI)	
	MAE	Theil coefficient	MAE	Theil coefficient
1 Quarter	2.72	1.09	1.36	0.31
2	1.90	0.87	1.58	0.41
4	1.45	0.73	1.95	0.60
8	1.32	0.80	2.73	0.69
Year 2 period	–	–	3.51	1.19
12	1.39	0.90	–	–

Source: Melliss and Whittaker (1998).

2.2.2.2. Summary. The results for the UK are similar to the findings relating to the US predictions. In fact, Burns (1986) provides a convenient comparison of US and UK forecasters. Burns concluded that there was little difference in forecasting performance in the two countries (a controversial issue at the time he drew this conclusion when the UK Treasury was under attack). Daub (1987) when comparing Canadian with US forecasting practice drew the same conclusion, explaining the similarities by the fact that innovations in forecasting practice have occurred because of an "efficient-like market in forecasting technology". However, Öller and Barot (2000) demonstrate that the quality of the performance in forecasting GDP across 13 European OECD countries is very variable relative to fluctuations in the economy despite this common forecasting technology.

3. Judging the forecasts

Forecasts might be judged in a number of ways. First, did the users find them useful? A more conventional criterion, based on quantitative error measures is to determine whether the forecasts have become more accurate over time and whether they were better than alternative predictions that would have been available to the users.

3.1. Qualitative results

Two surveys of the UK business community (in 1987 and 1994) revealed that macroeconomic forecasts were used for a variety of purposes. The respondents wanted forecasts about economic growth if they were concerned with real variables and predictions about monetary variables when concerned about financial decisions. These surveys also examined the perceived adequacy of the forecasting services and the criteria for selecting a service. The overwhelming impression from the 1994 survey (Fildes, 1995) was that macroeconomic forecasts were valuable to the companies, because they provided a framework within which corporate planning could take place. While accuracy was important, the availability of forecasts of the particular variable of interest for the appropriate time horizon was dominant, an observation also made by Canadian respondents (Daub, 1987).

McCloskey (1992) has expressed a critical view of the value of economic forecasts summarized as 'if you're so smart, why aren't you rich?'. She argues that if predictions had value, the forecasters would utilize them in profit making opportunities without resorting to intermediaries. This criticism does not apply to GDP or inflation forecasters because there are few opportunities to directly benefit from those predictions. It is more applicable to interest or exchange rate forecasts (see e.g. Leitch and Tanner, 1995).

While the business community has apparently felt well served by the forecasters, governments have been less pleased, often blaming their policy mistakes on forecast errors. However, Cairncross (1969) and Burns (1986), who were economic advisors to the UK government, note that economic forecasts are just one of the determinants of government policy. Even if not totally accurate, the forecasts create value because they offer a framework for sharing knowledge between the forecasters and the policy makers. The predictions offer a forum for consistent discussion that leads to "a new mental picture of how the world is likely to change" (Den Butter and Morgan, 1998).

In our view, this qualitative evidence shows that the forecasts are both valued and valuable.

3.2. Quantitative results

In using quantitative measures to judge the quality of the forecasts, we determine whether the accuracy of the forecasts has improved over time and whether the forecasts compare favorably with predictions obtained from naïve standards and/or time series benchmarks. We also examine the rationality of the forecasts.

3.2.1. Have the forecasts improved over time?

The past forty years have seen rapid growth in research in both macroeconomic theory and econometric methodology. We would expect therefore to see improvements in forecasting accuracy. The results from studies of this question have been contradictory. An early analysis of the ASA/NBER surveys indicated that accuracy had improved over time (Zarnowitz, 1979). More recent evidence for a longer time period is not so clear cut (Zarnowitz and Braun, 1992). However, this comparison had not been standardized for the volatility of the variables in the two periods. The Michigan Model's annual forecasts of real GNP over a 35-year period were baselined by calculating the ratio of the RMSEs to the standard deviation of the actual changes. These ratios declined monotonically from the 1950s to the 1980s, and McNees (1988) concluded that the accuracy of the real GNP forecasts had improved. [10]

The UK studies also provide mixed evidence about this question. Using standardized MAEs, Burns (1986) concluded that there had been an improvement in accuracy while Melliss (1997) noted that in the late 80s and early 90s there were signs of

[10] The same ratios that were used for the real GNP forecasts also decline for both the nominal GNP and inflation predictions.

deterioration. Similarly, the conclusions of Britton and Pain (1992) are mixed, but the errors had not been standardized. Öller and Barot's (2000) analysis of 13 European counties found no evidence of improvements in accuracy when they standardized the error statistics.

We conclude that the evidence does not show the forecasts have improved over time. The problem may be that the descriptive statistics, used to measure the difficulty of forecasting over different time periods, may not capture the difficulty of preparing forecasts in *any specific time period.* A more appropriate approach is to consider performance relative to a suitable benchmark that has the potential to eliminate idiosyncratic effects arising from the peculiarities of the particular period under study (Fildes, 1992).

3.2.2. Comparison with naïve standard; U coefficients

Theil's U coefficient is used to compare a set of forecasts against a naïve standard. If it is less than 1 in value, the forecasts have at least met the minimum standard of not being inferior to naïve predictions. The results are in general agreement: all of the US forecasts have U coefficients that are less than one. Some of the Us are, however, close to or larger than 0.9 (McNees and Ries, 1983; Zarnowitz, 1992). Moreover, U is only a descriptive statistic.

The UK Treasury's forecasts for GDP growth had Theil's $U < 1$ for forecast horizons as far ahead as 2 years (averaging 0.60 for the 12 quarters ahead forecasts and remaining robust across sub-periods for growth, Melliss, 1997). Relative performance of forecasts of GDP growth over the period improved as the forecasting horizon lengthened up to a year ahead. For inflation, Melliss' study yielded Us that were <1 only up to one year ahead. There was less consistency in the longer-term inflation forecasts, because the consistent underprediction in periods of increasing inflation, 1971–79, and 1985–89, led to poor longer term performance. Similar results were obtained for the NIESR forecasts of GDP and inflation (Britton and Pain, 1992) and for four macroforecasting groups (Holden and Thompson (1997).

Even though most forecasts meet a minimum standard, namely that the U coefficient be less than 1, we cannot conclude that any particular set of forecasts is significantly better than those obtained from the naïve models.

3.3. Comparison with time series benchmarks

The naïve models are not very stringent standards against which to compare forecasts. Time series models, such as ARIMA, VARS and Bayesian VARS (BVARs), are relatively simple to construct, can be used to generate predictions, and have been used as benchmarks to judge the accuracy of economic predictions.

3.3.1. US forecasts

Nelson (1972) compared the forecasts of an ARIMA model with those of a major model and found that the model was not markedly more accurate than the ARIMA benchmark. He then began to prepare ex ante real time predictions using ARIMA models. These forecasts are now generated by Fred Joutz and serve as a benchmark with which other predictions are compared (Nelson, 1984; Joutz, 1988; McNees, 1992a). Nelson and Joutz showed that from 1976.2 to 1985.4 the structural forecasts of growth rates and inflation were more accurate, at the shorter horizons, than the ARIMA predictions. As the horizon lengthened to three and four quarters, the Benchmark predictions were among the most accurate. This result was explained as possible poor dynamic specifications of the structural models that were corrected in the short run projections by judgmental adjustments. An additional explanation is that the structural models required forecasts of exogenous variables that may have been difficult to predict.

However, McNees' (1992a) results showed that from 1986.1 through 1991.3, structural models made more accurate forecasts of the real GNP growth rate than did the Benchmark alternative. Moreover, the relative accuracy of the structural models increased as the time horizon lengthened. On the other hand, the Benchmark's forecasts of the implicit GNP deflator were more accurate than about half of the model-based forecasts. The Nelson and Joutz results and McNees' findings apparently conflict. The discrepancies may be due to either the different time periods or the different methods that were used to evaluate the forecasts. A more important explanation for the difference in the results is that, over this period, econometricians were paying more attention to the dynamics of their models (Wallis, 1989; Diebold, 1998). This should have led model-based forecasts to improve relative to the Benchmark. The Benchmark forecasts have also been found comparable to the primarily judgmental ASA/NBER predictions (Zarnowitz and Braun, 1992).

Forecasts may also be compared with the predictions obtained from VARs or BVARs. Lupoletti and Webb (1986) compared the ex post forecasts of a VAR with the ex ante forecasts of three forecasting services for the period 1970–83. Over all horizons, all the forecasting services predicted real GNP better than did the VAR. The advantage was greatest at the short horizons, which is particularly significant because the VARs had a data advantage over the forecasting services. The VAR had been estimated via rolling regressions using after-the-fact revised data that were not available in real time. All of the professional forecasters also predicted inflation more accurately one quarter ahead and made comparable errors for the longer horizons. In fact, Webb (1995) showed that inflation forecasts from VARs were significantly worse than naïve predictions.

The accuracy of ex ante forecasts obtained from structural models and the predictions obtained from one BVAR were compared for the period 1980–85 by both Litterman (1986); McNees (1986) with similar results. The real GNP forecasts of the structural models were more accurate for the one-quarter ahead horizon, but thereafter the BVAR predictions made substantially smaller errors with the relative difference increasing with the length of the forecast horizon. On the other hand, the inflation predictions of the structural models were substantially superior to those of the BVAR with the relative performance of the structural models improving the longer the horizon.

On the basis of these studies we conclude that for the US, time series models provide competitive benchmarks for forecast evaluation and encompassing tests, sometimes outperforming more complex structural alternatives.

Table 4
The ratio of G-7 countries for which the BVAR forecasts were superior to the IMF's, 1980–87

Variable\model	Current year/one-step ahead		Year ahead/two-step-ahead	
	BVAR	BVAR+	BVAR	BVAR+
Growth	1/7	2/7	2/7	3/7
Inflation	2/7	5/7	3/7	3/7

NB: The BVAR+ forecasts include an extra quarter of data.

3.3.2. UK forecasts and other evidence

Early comparisons in the UK showed that the performance of ARIMA models was comparable to that of the structural models of the 1970s and provided prima facie evidence that the London Business School Model was at that time dynamically mis-specified (Longbottom and Holly, 1985). More recently, Wallis (1989), Holden and Broomhead (1990) and Holden (1997) have compared the forecasts of VARs with those of various modeling services. Wallis indicated that the forecasts of the VARs were not clearly superior to those obtained from structural models. However, the benchmark has to be carefully specified because BVARs generally outperform their unconstrained alternatives and some approaches to specification are better than others (Allen and Fildes, 2001). In a comparison with structural models Holden and Broomhead (1990) added supplementary evidence showing that (constrained) BVARs outperformed their unconstrained alternatives. Even using the BVARS, Holden (1997) found few instances where they outperformed their econometric competitors. [11]

In addition to the US and UK evaluations, time series methods have been compared with the forecasts of the IMF and OECD. Artis and Zhang (1990) compared BVAR forecasts for the G-7 group of countries with those produced by the IMF in the World Economic Outlook (WEO). The superior performance of the WEO is partially attributable to knowledge of current economic activity as Table 4 shows because the BVAR's performance improved when it used an additional quarter's information.

Zellner and his colleagues have extended the concept of what constitutes a benchmark model by (1) adding additional variables to an autoregressive model, (2) estimating the models over a cross-section of OECD countries, and (3) using time varying parameter models (Garcia-Ferrer et al., 1987; Zellner and Hong, 1989). All these extensions improved the fit and accuracy of the benchmark time series model. When the one-year ahead OECD forecasts were compared with the predictions of this final model, the time series model was superior.

These comparisons of time series predictions with ex ante forecasts are difficult to interpret. The forecaster has information about current quarter conditions that can be built into the predictions. This information would not be available to the time se-

[11] However, combining BVAR forecasts with the econometric forecasts showed they contained information (presumably their dynamic structure) that was not in the structural model forecasts.

ries model. On the other hand, the time series model is based on revised data that would not have been available to forecasters in real time.

A comparison of the UK and international forecasts with time series predictions indicates that time series models provide a rigorous benchmark for evaluating forecasts. This finding is in accord with US results.

3.4. Summary of the results

The quantitative findings indicated that all of the model-based forecasts outperformed the simplest naïve models. The predictive accuracy of time series models also approximated that of (1) the structural models and (2) the ASA/NBER surveys. Therefore, the time series models serve as useful benchmarks for judging other types of forecasts. However, our survey of the published evaluations indicates a serious analytical flaw: there are few tests for the statistical significance of the results. Nevertheless, the cumulative weight of evidence from the large number of comparisons we report here, across different data sets, different models and different countries, leads us to believe our conclusion is robust. Macroforecasts are also useful inputs into disaggregated forecasts. Business planners value them and empirical evidence suggests that they improve disaggregate forecasting accuracy (Ashley, 1983, 1988; Allen and Fildes, 2001).

4. Could the forecasts have been improved? Rationality tests

The naïve and time series comparisons merely provide some benchmarks for judging a set of predictions. Unfortunately, most evaluations have not provided tests for the statistical significance of the results. Could the forecasts have been improved? If ex post rationality studies show that the forecasts were biased or inefficient then, *in principle*, the accuracy of the ex ante predictions could have been significantly improved if this information had been taken into account. [12]

The results obtained from these rationality studies depend on a number of factors. These include (1) whether the analysis was performed on the "consensus" forecast obtained from a survey or from the individual forecasters who constituted the cross-section, (2) the date that the studies were undertaken, and (3) the econometric methods that were used in the analysis. In the US, there have been major studies of the rationality of the forecasts contained in the ASA/NBER, Blue Chip, Livingston and Michigan surveys. These analyses were conducted by Batchelor and Dua (1991), Davies and Lahiri (1995, 1999), Keane and Runkle (1990) and Zarnowitz (1985). In addition to these analyses of the survey forecasts, McNees (1978) and Joutz (1988) tested the forecasts of particular econometric models for rationality.

Studies of the rationality of UK forecasts have primarily examined econometric model based forecasts, (see for example, Mills and Pepper's (1999) analysis of model based forecasts—from LBS, NIESR and the UK Treasury). In addition, a recent

[12] This would not necessarily be possible in practice, as the optimal weights would be unknown.

study by Egginton (1999) has analyzed the predictions of various City of London forecasters.

4.1. Results: Inflation

4.1.1. US

In the US studies in which aggregated or consensus forecasts were examined, the hypothesis that the inflation predictions were rational was rejected in most instances. Rich's (1989) analysis of the Michigan data and the Batchelor–Dua study of the consensus Blue Chip forecasts are exceptions. The former study found that the mean forecast did not violate the unbiasedness, efficiency, and orthogonality conditions. The latter were only unbiased.

Zarnowitz (1985), using disaggregated forecasts, rejected the rationality hypothesis for the inflation predictions of *many* of the individuals in the ASA/NBER survey. Keane and Runkle (1990), however, showed that a smaller number were not rational. The findings differed because the estimation techniques and the data, that measured the outcomes, were not the same in the two studies. In turn, Davies and Lahiri (1999) have challenged the Keane and Runkle methodology and results and concluded that two thirds of the forecasters in their ASA/NBER sample failed the rationality test.

The participants in the Blue Chip survey made unbiased inflation forecasts but most failed one or more of the other rationality tests (Batchelor and Dua, 1991). In particular, the individuals failed to take account of published information and of the information contained in their *own* forecasts of other variables. These forecasters did worse in predicting inflation than their record in forecasting real growth. Batchelor and Dua provide two possible explanations of their results. First, 'irrational' forecasters tended to place more importance upon non-traditional economic theories than did the 'rational' forecasters. They also placed greater reliance upon econometric models and assigned lower weights to judgmental adjustments. While Batchelor and Dua concluded that the Blue Chip forecasts were unbiased, Davies and Lahiri (1999), using an improved methodology, found that more than half of the Blue Chip forecasters showed significant bias.

The analyses of the inflation forecasts of individual econometric models also yield conflicting results. McNees (1978) showed that some of the one-quarter ahead forecasts of inflation are unbiased but that most of the multi-period forecasts fail this test. Moreover, the results also depend on the econometric technique used. On the other hand, Joutz (1988) found that, for a longer time period, all the econometric models in his sample produce unbiased inflation predictions. However, Joutz also showed that available information was not always used efficiently. In particular, the inflation forecasts could have been improved if variables such as the degree of capacity utilization and labor market conditions had been more fully specified. [13]

[13] Just as the vintage of the actual data affected the Zarnowitz and Keane–Runkle results, mentioned above, the models in this sample have a smaller number of rejections of this rationality test if revised rather than preliminary data are used. For the UK Britton and Pain (1992, pp. 18–19) found data revisions did not have a significant influence on the results.

Similarly, Joutz and Stekler (2000) found that some of the inflation forecasts of the Federal Reserve were inefficient because there was evidence of autocorrelations in the residuals.

4.1.2. UK and other evidence

The UK evidence is broadly supportive of unbiasedness (Holden and Thompson, 1997; Melliss, 1997; Mills and Pepper, 1999), however, for some sub-periods it is rejected. In certain periods both the National Institute (Britton and Pain, 1992) and the Treasury (Melliss, 1997; Melliss and Whittaker, 1998) made systematic errors. Melliss also tested for strong form informational efficiency and found variables that could have improved forecasts in particular sub-periods for different forecast horizons, but no consistent effects were observed. Specifically, Melliss (1997) showed that longer-term forecasts of inflation could have been improved by including a capacity utilization variable. (This is similar to Joutz' results for the US.) Although individual short-term City forecasters were both biased and inefficient, these anomalies were small (Egginton, 1999). Melliss and Whittaker (1998) find evidence of irrationality because the successive updating of fixed event forecasts were inadequately revised.

There is only limited evidence about the rationality of other nations' forecasts. Kirchgassner (1993) examined the 6 month and 2nd half-year ahead predictions of growth, inflation and other macroeconomic indicators made by various German Research Institutes. The rationality hypothesis is not rejected for the 6-month predictions, but was consistently rejected for the 2 step ahead predictions. Ash et al. (1990) examined the rationality of OECD forecasts for the G-7 members. The inflation forecasts for Japan, Germany and France could have been improved by using additional available information. In addition, the EU record has been examined by Keereman (1999) where the forecasts were found to be mainly unbiased and efficient.

4.2. Results: Growth forecasts

4.2.1. US

The evidence regarding the rationality of the US real GNP predictions is also mixed. Most of the individuals in the NBER/ASA surveys made rational forecasts (Zarnowitz, 1985). The predictions of only 7 of 19 participants of the Blue Chip survey passed all of the rationality tests, [14] but the consensus forecast failed only one of the tests (Batchelor and Dua, 1990). McNees' (1978) study of three econometric models yielded mixed results (similar to those associated with the inflation forecasts). Joutz (1988) showed that unbiasedness was rejected for only two of six models, but that information was often used inefficiently. The use of monetary data could have improved these forecasts. On the other hand, Joutz and Stekler (2000) did not reject weak form efficiency for the real growth rate forecasts of the Federal Reserve.

[14] The improved methodology of Davies and Lahiri (1995) found that the forecasts of half of the Blue Chip participants were biased.

4.2.2. UK and other evidence

Bias is not a particularly common feature of the growth forecasts (Melliss (1997) analyzing the UK Treasury 1–8 quarters ahead; Holden and Thompson (1997) analyzing four macroeconomic modeling institutes 0–4 years ahead). However, Melliss's (1997) evidence about the efficiency of the growth forecasts is mixed (as it was with inflation). There is no consistent pattern, and the results are period specific, but "forecast inefficiency has been a common feature". This inefficiency can be attributed to dynamic misspecification because the inclusion of the lagged rate of growth was significant.

The OECD forecasts for the G-7 members were "relatively free of bias", but they often failed efficiency and consistency tests (Ash et al., 1990). In contrast to inflation, the errors in growth forecasts were non-systematic. The one year ahead growth forecasts made for 13 European countries showed very little evidence of either bias or weak form inefficiency Öller and Barot (2000), a result confirmed for the EU by Keereman (1999). In contrast, Loungani (2000) examined the *Consensus Forecasts* for 67 countries using both forecasts and forecast revisions (for a fixed outcome) and found weak form inefficiency.

4.3. Conclusions regarding rationality

These studies show that a number of individuals and models have generated rational forecasts but also demonstrate that other forecasters or models were not able to do this, perhaps acknowledging conflicting objectives such as the reputational value of differing from the consensus [15] (Batchelor and Dua, 1990; Spencer and Huston, 1993). Performance over different sub-periods (and different economic conditions) is, however, inconsistent. This finding holds across all countries.

Since the results are in conflict, meta-analysis techniques can be used to determine whether a definitive conclusion can be obtained by *combining all* of the evidence. A meta-analysis about the findings of US rationality studies (Goldfarb and Stekler, 2001) shows that the overall conclusions varied from year to year and with the time horizon of the forecast. In fact, the overall conclusions were sometimes reversed when an additional study was added. Performance is also country specific with North American forecasts more likely to be rational than in other parts of the world including the remainder of OECD. For the UK, growth and inflation are both "poorly forecast with both serious bouts of inflation and deep recession underpredicted" (Mills and Pepper, 1999).

The use of rationality tests, especially those that relate the forecast errors to known information, can thus best be viewed as important diagnostic checks to determine why the errors occurred and to improve the forecasting process and the quality of subsequent predictions.

[15] Burrell and Hall (1994) argue that UK City forecasters should be more likely than academic forecasters to recognize the possible benefits of differing from the consensus but fail to find any support for this in their data.

5. Choosing between forecasters (or methods)—or a combination

Our discussion has suggested that most quantitative forecasting methods yield similar results. Nevertheless, a number of studies have explicitly asked whether there is a best model, method or forecaster. Zarnowitz and Braun (1992) indicate that, in the US, the forecasting performance of any one individual relative to another is highly variable over time. A similar conclusion holds for the UK (Holden and Peel, 1986; Holden and Thompson, 1997). The clear implication is that there is no consistent best performer (Britton and Pain, 1992; McNees, 1992c; Burrell and Hall, 1994). It may be more appropriate to ask an alternative question: Are there some forecasters who are better or worse on average (Stekler, 1987; Batchelor, 1990)? McNees (1992c) asserts that while there is no forecaster or method, which is best all of the time, there are some forecasting services that consistently performed poorly.

Given that there is no "best" forecasting approach, it is natural to determine whether a combination of forecasts might produce better results. The combined forecast is of the form

$$Y_t^* = \alpha_t + \sum_i \beta_{it} \widehat{Y}_{it}$$

where Y_t^* is the combined forecast of the variable obtained from the i individual sources. While considerable research has been undertaken to determine how best to choose the coefficients, α_t and β_{it}, the evidence suggests that the simple approach of averaging the individual predictions works well (Lupoletti and Webb, 1986; Clemen and Winkler, 1986; Clemen, 1989).

In most cases, combining predictions reduces the out-of-sample errors. Although the seminal article on forecast combination is that by Bates and Granger (1969), Zarnowitz (1967) had earlier noted that the average of several GNP forecasts was better than the individual predictions. Similarly, Nelson (1972) and Cooper and Nelson (1975) showed that when econometric forecasts were combined with ARIMA estimates, the combined forecast had smaller errors than were obtained from the models alone. More recent evidence also suggests that combining typically leads to improved accuracy (Lupoletti and Webb, 1986; Clemen and Winkler, 1986). Clemen's (1989) review cites additional studies that confirm these results. The benefits of combining are greatest when forecasts that are generated from different methods or theories are combined (Batchelor and Dua, 1995). In the same vein, Graham (1996) pointed out that when the forecasts of economists are correlated, the effective number of independent forecasts is reduced.

If ex post, studies show that the best of the model-based forecasts is more accurate than the combination, some may question the effectiveness of combining (Holden and Thompson, 1997). However, the fact that a *particular* model is more accurate than the combination is not known a priori. A complementary view is to see the combination of forecasts as a test of the specification of the alternative models. Proponents argue that the 'true' model should be able to explain the results of other models. Encompassing tests would indicate whether or not there was information in one model's forecast that was not in the other's (see Chong and Hendry, 1986;

Diebold, 1989). This test may be performed by regressing the actual values on the forecasts of the two models, i.e.

$$Y_t = \alpha + \beta_1 \widehat{Y}_{1t} + \beta_2 \widehat{Y}_{2t}$$

and then determining whether the vector of coefficients $(\alpha, \beta_1, \beta_2)$ is $(0, 1, 0)$, in which case Model 1 encompasses Model 2, or whether it is $(0, 0, 1)$, in which case the second encompasses the first. If any other result is obtained, there is useful information in both models and both models are mis-specified. [16] (This test is similar to the conditional efficiency rationality test where a second forecast is included in a rationality test for efficiency to see if it adds any information.) Hence, *some model builders advocate that more effort should be devoted to obtaining the true model specification rather than combining the forecasts—although current users of the forecasts would no doubt prefer improved accuracy.*

6. The role of judgment and forecast error decomposition

The role of judgment in the forecasting process has long been a subject of analysis. In the 1960s the question of interest was whether judgmental or model based forecasts were more accurate. This question is no longer relevant, for forecasting methods are rarely used in their pure form. All individuals use at least some quantitative techniques, and most econometricians make judgmental adjustments to their model-based forecasts. [17]

Models may be adjusted for a number of related reasons: (1) structural breaks, where changes have occurred in a particular equation that post-date the sample data on which the model was estimated; (2) specific parameter estimation errors, where part of a recently observed residual may be due to a change in a parameter during the postestimation sample, for example a change in the institutional framework within which wages are determined, e.g. an incomes policy which leads to a constraint on a parameter in the model; (3) measurement errors, which typically arise when data used in the model-based forecasts are revised, and (4) current information derived from sources outside the model that gives an insight into possible inadequacies in the model-based forecast.

The residuals of particular equations may be modified judgmentally to take account of any of these factors. In practice, either mechanical rules or subjective estimates are used to modify the equations (Young, 1982; Turner, 1990). Todd (1992) argues that when adjustments are made, the reasons for doing so should be stated explicitly, and Moss et al. (1994) have even proposed the use of an expert systems approach that codifies the basis for making such adjustments to the model. In this way, ex post evaluations would be able to determine why the model-based forecast was adjusted.

[16] See Harvey et al. (1998) for a test that is robust to non-normality, a common problem.

[17] Ray Fair does not adjust the forecasts of his model and is critical of this procedure.

In order to determine the effectiveness of these judgmental adjustments to model-based forecasts, it is first necessary to decompose the ex ante forecast error into various sources. The forecast f depends on the econometric specification, the exogenous variables, X, the corresponding forecasts of these variables, $\widehat{X}$, and any adjustment, Ad, the forecaster might make to the model output. This relationship is shown in the following decomposition:

	Model error (ey)	(a) $\mathrm{ey} = Y - f(X, 0)$
Less	Contribution of residual adjustment	(b) $-\{f(X, \mathrm{Ad}) - f(X, 0)\}$
Plus	Contribution of exogenous variable errors	(c) $+\{f(X, \mathrm{Ad}) - f(\widehat{X}, \mathrm{Ad})\}$
=	Forecast error	(d) $Y - f(\widehat{X}, \mathrm{Ad})$

If there were no exogenous variables in the system, it would be possible to determine the value of judgment merely by examining Eq. (b), the difference between the ex ante forecasts, $f(\widehat{X}, \mathrm{Ad})$ which incorporated judgmental adjustments, and the actual ex post unadjusted model projections, $f(X, 0)$. However, the effects of predicting the exogenous variables also make it more difficult to evaluate the role of judgment in model based forecasts. [18] We thus conclude that in any evaluation of ex ante forecasting accuracy, in the typical case where the forecaster has used judgment in model building, the other errors cannot be uniquely distinguished from the judgmental effects. [19]

This indeterminacy occurs, for example, when econometricians make assumptions about the exogenous variables rather than adjustments to subjectively influence their forecasts. This finding has been observed in studies in which the ex post actual values of the exogenous variables have been substituted for the ex ante estimates. This (together with the residual adjustments) explains the often quoted conclusion that ex ante macroforecasts are more accurate than the corresponding ex post forecasts.

In forecasting applications, where the aim is improved accuracy, it is hard to argue against subjective intervention as a matter of principle. Rather, the issue is whether the adjustments improve on the mechanical forecasts. Fair (1984), Zarnowitz (1992, pp. 408–410) and McNees (1982, 1990) have concluded that the evidence is weakly in their favor, but the adjustments were not optimal. As McNees (1990) notes, the prevailing tendency is “to place too much weight on the specific circumstances and too little on the model”. [20] In other words, the forecasters

[18] To take these effects into account Wallis and Whitley (1991) proposed that the ex post mechanical forecasts derived from the actual values of the exogenous variables be compared with mechanical forecasts based on the assumed values of the exogenous variables used in the ex ante predictions, i.e. $f(X, 0) - f(\widehat{X}, 0)$.

[19] The problem becomes even more complex if, as Hendry (1997) proposes, the model error decomposition incorporates various mis-specifications including parameter and intercept mis-specification over the forecast period.

[20] Donihue’s (1993) analysis of the Michigan model yielded the same result. Such a conclusion can also be found in the more general forecasting literature (Armstrong and Collopy, 1998).

overcompensated. In the UK, Wallis (1986, 1987, 1989) analyzed their effectiveness and concluded, "the adjustments reduce the size of forecast errors". The general picture is one of 'value-added by forecasters to the accuracy of forecasts'. *In summary, any evaluation of a macromodel's forecasts cannot easily be disentangled from the judgment of the people running the model.*

7. General summary and suggestions for improving forecasting

Based on this survey of the published results of forecast evaluations, what can we conclude about the state of macroeconomic forecasting?

7.1. General summary

We have found that most forecasters fail to predict recessions in advance and sometimes fail to recognize them contemporaneously. Forecasters also seem to make systematic errors such as underestimating growth during periods of economic expansion, overestimating it during declines, underpredicting inflation when it is accelerating and overpredicting when it is decelerating.

Despite the relatively large quantitative errors observed, almost all forecasts are superior to the predictions that could have been obtained from naïve models and often better than those generated by time series models. In most comparisons, however, the results had not been tested for statistical significance, and we can not determine whether one set of forecasts is significantly better than another ex post, nor can we identify ex ante the future best performers. For significant periods, particularly for inflation, the forecasts are 'irrational' and could be improved by the efficient use of available information. In addition, there is little evidence that the forecasts have improved over time.

There is one result about which there is general agreement, namely no one forecasting method or one model or one individual does best all of the time. This has led to an effective suggestion for improving the quality of the forecasts: combine the predictions of different individuals or methods. We also found that there was agreement that the use of judgmental adjustments improved the quality of model forecasts. This accords with the generally accepted practices of model builders.

Both theoretical and empirical studies have shown economic forecasting to be potentially valuable (and market arguments show that it is of value to organizations). However, this leaves open the question of the value of specific macroforecasts of GNP. After reporting that even poor forecasts of interest rates (relative to naïve competitors) had value through various trading rules (Leitch and Tanner, 1995) they go on to argue that directional GNP forecasts are similarly valuable, a point supported by Öller and Barot (2000). Unfortunately this important field of study has received too little attention from the academic community apart from financial valuation models. which directly integrate ex ante forecasting.

7.2. Improving accuracy

We next report on efforts that have already been undertaken to improve forecast accuracy. First, the major modeling failures of the 1970s led to various theoretical enhancements of the models to include supply side variables. Some forecast failures, for example the recession of 1979–81 in the UK were explainable ex post by existing models. Other events, such as the fall of oil prices, required a reappraisal of the price response function (Wallis, 1987). Forward-looking expectations were incorporated into most macromodels because of the Lucas critique and the notion of rational expectations.

In addition new models have been developed that provide a rigorous theoretical basis for macroeconomic analysis. These models are based on rational choices made by individual economic agents facing alternative technologies. Diebold (1998) believes that this 'dynamic stochastic equilibrium modeling' is a 'marvelous union of modern macroeconomic theory and non-structural time-series econometrics' that will improve forecasting performance. In any event, such a theoretical approach leaves forecasters less susceptible to the criticism that their models are ad hoc.

These theoretical developments have been matched by statistical innovations. The use of time series benchmarks had pointed out some inadequacies of structural models, especially their dynamic specifications. The comparison of model forecasts with those generated by time series thus led to improved model specifications as well as a more refined evaluation methodology based on co-integration. A greater use of economic indicators to improve predictive performance has also been suggested. For example, Zellner and Hong (1989) analyze country growth through a constructed indicator of world growth and Stock and Watson (1999) model US inflation using an indicator based on 168 disaggregate activity indicators. The latest developments by Zellner and Chen (2001) combine both aggregate and disaggregate indicators, showing the benefits of including both, through Bayesian shrinkage procedures. The evaluations of these approaches with established methods have used revised rather than real-time data. *Given this problem we cannot yet conclude that the new approaches lead to improved predictive accuracy.*

The forecasting failures that we have documented may also arise from non-linearities (Ormerod in Granger et al., 1998). The problem is that data observed in one part of a non-linear system is not informative as to behavior elsewhere (Ormerod and Campbell, 1997). On the other hand, Swanson and White (1997) show that, in general, a flexible specification (where parameters and variables are permitted to enter and exit from a single equation VAR) is more helpful than the inclusion of non-linear effects through neural networks. Similarly, Tkacz (2001) had little success using neural network methods although they proved better in longer term forecasts of Canadian GDP. Qi (2001) claims greater success when forecasting the probability of a US recession. Non-linear models have also been explored because they are a natural extension of the benchmark models that we have described and because stylized macroeconomic facts indicate that business cycle expansions and contractions differ. This has led to the idea of using regime-switching models. (Diebold (1998) has a fuller discussion; see Ivanova et al. (2000), Bidarkota (2001) for examples.) However, the importance of non-linear (perhaps chaotic) processes for forecasting is in dispute. Diebold (1998) points out that only in financial markets have they proved successful while Clements and Hendry (1998a) argue that well-behaved near linear models can be found equivalent to the non-linear representations, and these simpler models are more suitable for forecasting.

The effect of these innovations should have been improved performance (relative to some suitable benchmark). Diebold (1998) has taken a sanguine view of the success of the modeling effort, describing the developments in economic theory and time series econometrics as a successful tradition. His ''future of macroeconomic forecasting'' is much like his past—more theoretical improvements derived from 'dynamic stochastic equilibrium models', joined with improvements in estimation for such models (rather than the current reliance on calibration). Whether this approach will be more successful than what has been observed in the past is an open question.

New methodologies that focus on improving forecast accuracy have also been suggested (Granger, 1995; Clements and Hendry, 1998a,b). These approaches abandon the premise that there is a 'correct' structural model that is stable over time. Hendry (1997) and Clements and Hendry (1998b) analyze the sources of quantitative forecast errors by assuming that econometric models are being used to generate those predictions and that a structural change occurs. They argue that the key sources of forecasting failure are deterministic shifts (in the level or trend equilibrium value) and not parameter instability, which is of lesser significance. An inadequate model can perform substantially better than the 'correct model' if the former is insensitive to any structural break. These studies show why structural breaks can cause models that provided a good fit in the sample to perform badly when generating ex ante forecasts beyond the period of fit. Among the econometric techniques identified for improving the accuracy of forecasts in the presence of structural breaks is the use of judgmental adjustments, a procedure long utilized by model builders and, that does indeed improve the accuracy of forecasts. The explanation is that the adjustment is in effect an 'intercept correction' that has the potential to correct for a structural break. [21]

This summary of current developments leads to some suggestions about additional research that should be undertaken to improve forecast accuracy. In the area of methodology we suggest:

(1) Develop a methodology for estimating parameter shifts [22] and an *ex post explanation of why such shifts and other errors occur.* Despite the profession's documentation that systematic forecast errors have occurred, we do not have

[21] In a world where such structural breaks are common in the forecast period, non-causal models will often perform well and the imposition of unit roots in conflict with the in-sample evidence improves accuracy as the model immediately corrects for the intercept shift.

[22] Clements and Hendry (1998b, Chapter 12) discuss various tests but it remains to be determined whether they can be used to help real-time forecasting. An example of the improvements from forecasting accuracy arising from the identification of structural breaks with past economic events is given by Junttila (2001) when forecasting Finnish inflation.

specific information about the particular forecast failures. For example, the errors made in forecasting the UK recession of the early 1990s are not explainable by current models (Britton and Pain, 1992). This suggests that these models have omitted variables and structural change parameters that remain to be identified.

(2) Develop ex ante approaches for identifying structural change and forecasting its effects. Linked with (1) above this would in part be a cliometric history of forecasting failure interpreted through existing and past macromodels. An example of such an approach is found in Wallis (1987).

To expand on these recommendations, if intercept adjustments (or revised estimates of other parameters in the model) are to be made, how will a forecaster know when to make them, i.e. how will the structural change or regime change be recognized in advance or even contemporaneously? The comparative analysis of differences in specific forecasts between alternative models, as described above, will sometimes be informative and lead to model improvement. But the 'core assumptions' that lead most forecasters to make the same types of mistake have not been investigated. *What is required is a thorough understanding of the intellectual or cognitive processes that forecasters use in making forecasts and adjusting their models.*

In order to obtain this understanding of forecasting processes, it would be necessary to build and test models of forecaster behavior. Forecasters have obviously adopted some heuristics; for example, the tendency to produce damped forecasts of inflation towards the mean and the lack of efficiency in inflation forecasts is reminiscent of the 'regression fallacy'. Some efforts have already been directed towards identifying such rules (e.g. Ehrbeck and Waldman, 1996; Lamont, 1995, 2002). For example Lamont (2002) found US forecasters became more radical (and less accurate) as they became older while Ashiya and Doi (2001), examining Japanese forecasters, observed no such effects. These models postulate that individuals might be motivated by factors other than accuracy when they prepare their forecasts. The perceptions of their clients and their response to the consensus forecast (Batchelor and Dua, 1990, 1991) and the need to avoid sudden adjustments (even if new information has just become available) lead to apparent economic irrationality in the forecasts. If forecasters are so motivated, further research is required to determine the arguments of their loss functions and whether the losses are asymmetric (as they typically are in financial markets). The institutional framework in which the forecasts are produced (e.g. academic, governmental, financial sector) as well as the human capital invested in the forecasting organization may also affect accuracy (Daub, 1987; Smith, 1994; Baimbridge and Whyman, 1997) and should be investigated.

Our third suggestion for further research has, therefore, a rather different focus. Performance can be improved by making more effective use of the modelers' expertise, codifying and debiasing their knowledge through:

(3) Conducting further research into the role of judgement and how best to incorporate additional information sources within the model's framework in order to overcome at least some of the problems posed by structural change. These might include variables such as leading indicators or consumer confidence although effects are clearly non-linear (Batchelor and Dua, 1998).

Throughout, the timeliness and accuracy of data has proved important in determining forecast accuracy. Our final point is to reiterate the importance of having accurate real-time data that reflect the current state of the economy to incorporate into the modeling process. Even now, consistent data revisions have led to persistent and unnecessarily large forecast errors (Öller and Barot, 2000).

In summary, this paper has argued that macroeconomic forecasts need to be improved—they serve a wide variety of purposes within both public and private sector organizations and the value of any improvements has been shown to be high. Unfortunately, as Smith (1994) has remarked, the industry structure in the UK is such as to lead to sub-optimal investment in research aimed at evaluating and improving the quality of the macroeconomic forecasting industry. The picture in the US (Klein, 1991; Wallis, 1993) is equally gloomy. While there is considerable research resource spent in developing macroeconomic theory and to a lesser extent, statistical models, little attention has been given to the evaluation and improvement of forecasting performance. [23] As a consequence a research program for improving such forecasts was suggested and *it is our fervent belief that success in any of these areas should also lead to an improved understanding of economic processes.*

Acknowledgements

The authors would like to thank Ken Holden and staff in the UK Treasury for help in data collection, also the many commentators on the original draft, many of whom had themselves made contributions to researching macroeconomic forecasting. Only through discussion and debate can progress be made.

References

Allen, P.G., Fildes, R., 2001. The principles of econometric forecasting. In: Armstrong, J.S. (Ed.), The Principles of Forecasting. Kluwer Academic Publishers, Norwell, MA, forthcoming.

Anderson, P.S., 1997. Forecast errors and financial developments. Working Paper No. 51, Bank For International Settlements, Basle, Switzerland.

Armstrong, J.S., Collopy, F., 1992. Error measures for generalizing about forecasting methods: Empirical comparisons. International Journal of Forecasting 8, 69–80.

Armstrong, J.S., Collopy, F., 1998. Integration of statistical methods and judgement for time series forecasting: Principles from empirical research. In: George, W., Goodwin, P. (Eds.), Forecasting with Judgement. Wiley, Chichester, UK, pp. 99–137.

Armstrong, J.S., Fildes, R., 1995. On the selection of error measures for comparisons among forecasting methods. Journal of Forecasting 14, 67–71.

Artis, M.J., Zhang, W., 1990. BVAR forecasts of the G-7. International Journal of Forecasting 6, 349–362.

[23] Interestingly, the UK Macroeconomic Modelling Bureau is not now funded to examine the forecasting record of the major UK service providers. The funding agency thought it would be of limited academic interest. The US macromodelers have found it particularly difficult to carry out convincing comparative studies (Wallis, 1993).

Ash, J.C.K., Smyth, D.J., Heravi, S.M., 1990. The accuracy of OECD forecasts of the international economy. International Journal of Forecasting 6, 379–392.

Ash, J.C.K., Smyth, D.J., Heravi, S.M., 1998. Are OECD forecasts rational and useful? A directional analysis. International Journal of Forecasting 14, 381–391.

Ashiya, M., Doi, T., 2001. Herd behavior and Japanese economists. Journal of Economic Behavior and Organization 46, 343–346.

Ashley, R., 1983. On the usefulness of macroeconomic forecasts as inputs to forecasting models. Journal of Forecasting 2, 211–223.

Ashley, R., 1988. On the optimal use of non-optimal forecasts. International Journal of Forecasting 4, 363–376.

Ashley, R., 1998. New technique for postsample model selection and validation. Journal of Economic Dynamics & Control 22, 647–665.

Baillie, R.T. et al., 1993. On the limitations of comparing mean square forecast errors: Commentary. Journal of Forecasting 12, 639–667.

Baimbridge, M., Whyman, P., 1997. Institutional macroeconomic forecasting performance of the UK. Applied Economic Letters 4, 373–376.

Barker, T., 1985. Forecasting the economic recession in the UK. 1979–1982: A comparison of the model-based ex ante forecasts. Journal of Forecasting 4, 133–151.

Batchelor, R., 1990. All forecasters are equal. Journal of Business and Economic Statistics 8, 143–144.

Batchelor, R., Dua, P., 1990. Forecaster ideology, forecasting method and the accuracy of economic forecasts. International Journal of Forecasting 6, 311–316.

Batchelor, R., Dua, P., 1991. Blue chip rationality tests. Journal of Money, Credit, and Banking 23, 692–705.

Batchelor, R., Dua, P., 1995. Forecaster diversity and the benefits of combining forecasts. Management Science 41, 68–74.

Batchelor, R., Dua, P., 1998. Improving macro-economic forecasts; the role of consumer confidence. International Journal of Forecasting 14, 71–81.

Bates, J.M., Granger, C.W.J., 1969. The combination of forecasts. Operational Research Quarterly 20, 451–468.

Bidarkota, P.V., 2001. Alternative regime switching models for forecasting inflation. Journal of Forecasting 20, 21–35.

Blix, M., Wadefjord, J., Ulricka, W., Martin, A., 2001. How good is the forecasting performance of major institutions. Penning - Och Valutapolitik 3, 37–67.

Bodkin, R., Klein, L.R., Marwah, K., 1991. A History of Macroeconomic Model Building. Edward Elgar, Aldershot, UK.

Bolger, F., Harvey, N., 1998. Heuristics and biases in judgemental forecasting. In: George, W., Goodwin, P. (Eds.), Forecasting with Judgement. Wiley, Chichester, UK, pp. 99–137.

Britton, A., Pain, N., 1992. Economic forecasting in Britain—Report series no. 4, National Institute of Economic and Social Research, London.

Burns, T., 1986. The interpretation and use of economic predictions. Proceedings of the Royal Society, Series A 407, 103–125.

Burrell, A., Hall, S., 1994. A comparison of macroeconomic forecasts. In: Economic Outlook. London Business School, London.

Cairncross, A., 1969. Economic forecasting. Economic Journal 79, 797–812.

Chatfield, C., 1993. Calculating interval forecasts. Journal of Business and Economic Statistics 11, 121–135.

Cho, D.W., 2002. Do revisions improve forecasts? International Journal of Forecasting 18, 107–115.

Chong, Y.Y., Hendry, D.F., 1986. Econometric evaluation of linear macro-econometric models. Review of Economic Studies 53, 671–690.

Clemen, R.T., 1989. Combining forecasts: A review and annotated bibliography. International Journal of Forecasting 5, 559–583.

Clemen, R.T., Winkler, R.L., 1986. Combining economic forecasts. Journal of Business and Economic Statistics 4, 39–46.

Clements, M.P., 1997. Evaluating the rationality of fixed-event forecasts. Journal of Forecasting 16, 225–239.

Clements, M.P., Hendry, D.F., 1993. On the limitations of comparing mean square forecast errors. Journal of Forecasting 12, 617–637.

Clements, M.P., Hendry, D.F., 1998a. Forecasting economic processes—a reply. International Journal of Forecasting 14, 139–143.

Clements, M.P., Hendry, D.F., 1998b. Forecasting Economic Time Series: The Marshall Lectures on Economic Forecasting. Cambridge University Press, Cambridge, UK.

Cooper, J.P., Nelson, C.R., 1975. The ex ante prediction performance of the St. Louis and FRB-MIT-Penn econometric models and some results on composite predictions. Journal of Money, Credit, and Banking 7, 1–32.

Corrado, C., Greene, M., 1988. Reducing uncertainty in short term predictions; linkage of monthly and quarterly models. Journal of Forecasting 7, 77–102.

Daub, M., 1987. Canadian Economic Forecasting: In a World where All's Unsure. McGill-Queens University Press, Kingston and Montreal, Canada.

Davies, A., Lahiri, K., 1995. A new framework for analyzing survey forecasts using three-dimensional panel data. Journal of Econometrics 68, 205–227.

Davies, A., Lahiri, K., 1999. Reexamining the rational expectations hypothesis using panel data on multiperiod forecasts. In: Hsiao, C. et al. (Eds.), Analysis of Panels and Limited Dependent Variable Models. Cambridge University Press, Cambridge, UK, pp. 226–254.

Den Butter, F.A.G., Morgan, M.S., 1998. What makes the models–policy interaction successful? Economic Modelling 15, 443–475.

Diebold, F.X., 1989. Forecast combination and encompassing: Reconciling two divergent literatures. International Journal of Forecasting 6, 589–592.

Diebold, F.X., 1998. The past, present and future of macroeconomic forecasting. Journal of Economic Perspectives 12, 175–192.

Diebold, F.X., Gunter, T.A., Tay, A.S., 1998. Evaluating density forecasts with applications to financial risk management. International Economic Review 39, 863–883.

Diebold, F.X., Mariano, R., 1995. Comparing predictive accuracy. Journal of Business and Economic Statistics 13, 253–263.

Diebold, F.X., Tay, A.S., Wallis, K.F., 1999. Evaluating density forecasts of inflation: The survey of professional forecasters. In: Engle, R., White, H. (Eds.), Festschrift in Honor of C.W.J. Granger. Oxford University Press, Oxford, pp. 76–90.

Donihue, M.R., 1993. Evaluating the role judgment plays in forecast accuracy. Journal of Forecasting 12, 81–92.

Egginton, D.M., 1999. Testing the efficiency and rationality of City forecasts. International Journal of Forecasting 15, 57–66.

Ehrbeck, T., Waldman, R., 1996. Why are professional forecasters biased? Agency versus behavioral explanations. Quarterly Journal of Economics 111, 21–40.

Fair, R.C., 1980. Estimating the expected predictive accuracy of econometric models. International Economic Review 21, 355–378.

Fair, R.C., 1984. Specification, estimation and analysis of macroeconometric models. Harvard University Press, Cambridge, MA.

Fildes, R., 1992. The evaluation of extrapolative forecasting methods with discussion. International Journal of Forecasting, 81–111.

Fildes, R. (Ed.), 1995. World Index of Economic Forecasts, fourth ed. Gower, Aldershot, Hampshire, UK.

Garcia-Ferrer, A., Highfield, R.A., Palm, F., Zellner, A., 1987. Macroeconomic forecasting using pooled international data. Journal of Business and Economic Statistics 5, 53–67.

Goldfarb, R., Stekler, H.O., 2001. Combining the results of rationality studies – What did we know and when did we know it. Indian Economic Review 36, 269–300.

Graham, J.R., 1996. Is a group of economists better than one? Than none? Journal of Business 69, 193–232.

Granger, C.W.J., 1995. Can we improve the perceived quality of economic forecasts. Journal of Applied Econometrics 11, 455–473.

Granger, C.W.J., Newbold, P., 1977. Forecasting Economic Time Series. Academic Press, San Francisco.

Granger, C.W.J., Ormerod, P., Smith, R., 1998. Comments on forecasting economic processes. International Journal of Forecasting 14, 133–137.

Harvey, D., Leybourne, S., Newbold, P., 1997. Testing the equality of prediction mean squared errors. International Journal of Forecasting 13, 281–291.

Harvey, D., Leybourne, S., Newbold, P., 1998. Tests for forecast encompassing. Journal of Business and Economic Statistics 16, 254–259.

Hatanaka, M., 1975. The underestimation of variations in the forecast series. International Economic Review 16, 151–160.

Hendry, D.F., 1997. The econometrics of macroeconomic forecasting. Economic Journal 107, 1330–1357.

Holden, K., 1997. A comparison of forecasts from UK economic models and some Bayesian vector autoregressive models. Journal of Economic Studies 24, 243–257.

Holden, K., 1995. Vector autoregression modelling and forecasting. Journal of Forecasting 14, 410–420.

Holden, K., Broomhead, A., 1990. An examination of vector autoregressive forecasts of the UK economy. International Journal of Forecasting 6, 11–23.

Holden, K., Peel, D.A., 1985. An evaluation of quarterly National Institute forecasts. Journal of Forecasting 4, 227–234.

Holden, K., Peel, D.A., 1986. An empirical investigation of combinations of economic forecasts. Journal of Forecasting 5, 229–242.

Holden, K., Peel, D.A., 1990. On testing for unbiasedness and efficiency of forecasts. Manchester School 58, 120–127.

Holden, K., Thompson, J., 1997. Combining forecasts, encompassing and the properties of UK macroeconomic forecasts. Applied Economics 29, 1447–1458.

Howrey, E.P., 1978. The use of preliminary data in econometric forecasting. Review of Economics and Statistics 60, 193–200.

Ivanova, D., Lahiri, K., Seitz, F., 2000. Interest rate spreads as predictors of German inflation and business cycles. International Journal of Forecasting 16, 39–58.

Jenkins, G.M., 1982. Some practical aspects of forecasting in organizations. Journal of Forecasting 1, 3–21.

Jeong, J., Maddala, G.S., 1991. Measurement errors and tests for rationality. Journal of Business and Economic Statistics 9, 431–439.

Joutz, F.L., 1988. Informational efficiency tests of quarterly macroeconometric forecasts from 1976 to 1985. Managerial and Decision Economics 9, 311–320.

Joutz, F., Stekler, H.O., 2000. An evaluation of predictions of the Federal Reserve. International Journal of Forecasting 16, 17–38.

Junttila, J., 2001. Structural breaks, ARIMA model and Finnish inflation forecasts. International Journal of Forecasting 17, 203–230.

Keane, M.P., Runkle, D.F., 1990. Testing the rationality of price forecasts: New evidence from panel data. American Economic Review 80, 714–735.

Keereman, F., 1999. The track record of the Commission Forecasts. Economic Papers Number 137, European Commission, Directorate-General for Economic and Financial Affairs, Brussels.

Kirchgassner, G., 1993. Testing weak rationality of forecasts with different time horizons. Journal of Forecasting 12, 541–558.

Klein, L.R. (Ed.), 1991. Comparative Performance of US Econometric Models. Oxford University Press, Oxford.

Kolb, R.A, Stekler, H.O., 1990. The lead and accuracy of macroeconomic forecasts. Journal of Macroeconomics 12, 111–123.

Lamont, O., 1995. Macroeconomic forecasts and microeconomic forecasters. NBER Working paper 5284, New York.

Lamont, O.A., 2002. Macroeconomic forecasts and microeconomic forecasters. Journal of Economics Behaviour and Organization 48, 265–280.

Leitch, G., Tanner, J.E., 1995. Professional economic forecasts: Are they worth their costs? Journal of Forecasting 14, 143–157.

Litterman, R.B., 1986. Forecasting with Bayesian vector autoregressions—five years of experience. Journal of Business and Economic Statistics 4, 25–38.

Longbottom, J.A., Holly, S., 1985. The role of time series analysis in the evaluation of econometric models. Journal of Forecasting 4, 75–87.

Loungani, P., 2000. How accurate are private sector forecasts? Cross-country evidence from Consensus Forecasts of output growth. IMF Working Paper WP/00/77, Internal Monetary Fund, Washington, DC.

Lupoletti, V.M., Webb, R.H., 1986. Defining and improving the accuracy of macroeconomic forecasts: Contributions from a VAR model. Journal of Business 59, 263–285.

McCloskey, D., 1992. The art of forecasting from ancient to modern times. Cato Journal 12, 23–48.

McNees, S.K., 1976a. An evaluation of economic forecasts: Extension and update. New England Economic Review, 30–44.

McNees, S.K., 1976b. The forecasting performance in the 1970s. New England Economic Review, 1–13.

McNees, S.K., 1978. The rationality of economic forecasts. American Economic Review and Proceedings 68, 301–317.

McNees, S.K., 1979. The forecasting record for the 1970s. New England Economic Review, 1–21.

McNees, S.K., 1982. The role of macroeconomic forecasting and policy analysis in the US. Journal of Forecasting 1, 37–48.

McNees, S.K., 1985. Which forecast should you use? New England Economic Review, 36–42.

McNees, S.K., 1986. Forecasting accuracy of alternative techniques: A comparison of US macroeconomic forecasts. Journal of Business and Economic Statistics, 4, 5, 15, 23.

McNees, S.K., 1987. Consensus forecasts: Tyranny of the majority. New England Economic Review, 15–21.

McNees, S.K., 1988. How accurate are macroeconomic forecasts? New England Economic Review, 15–26.

McNees, S.K., 1990. The role of judgment in macroeconomic forecasting accuracy. International Journal of Forecasting 6, 287–299.

McNees, S.K., 1992a. How large are economic forecast errors? New England Economic Review, 25–33.

McNees, S.K., 1992b. The 1990–91 recession in historical perspective. New England Economic Review (Jan–Feb), 3–22.

McNees, S.K., 1992c. The uses and abuses of 'consensus' forecasts. Journal of Forecasting 11, 703–710.

McNees, S.K., Ries, J., 1983. The track record of macroeconomic forecasts. New England Economic Review, 5–18.

Melliss, C., 1997. The Treasury forecasting record: An evaluation. ESRC Macroeconomic Modelling Bureau. University of Warwick, Coventry, UK.

Melliss, C., Whittaker, R., 1998. The Treasury forecasting record: Some new results. National Institute Economic Review 164, 65–79.

Mills, T.C., Pepper, G., 1999. Assessing the forecasters. International Journal of Forecasting 15, 247–257.

Mincer, J., Zarnowitz, V., 1969. The evaluation of economic forecasts. In: Mincer, J. (Ed.), Economic Forecasts and Expectations. National Bureau of Economic Research, New York.

Moore, G., 1969. Forecasting short-term economic change. Journal of the American Statistical Association 64, 1–22.

Moss, S., Artis, M., Ormerod, P., 1994. A smart automated macroeconometric forecasting system. Journal of Forecasting 13, 299–312.

Nelson, C.R., 1972. The prediction performance of the FRB-MIT Penn model of the US economy. American Economic Review 62, 902–917.

Nelson, C.R., 1984. A benchmark for the accuracy of econometric forecasts of GNP. Business Economics, 52–58.

Nordhaus, W.D., 1987. Forecasting efficiency: Concepts and applications. Review of Economics and Statistics 69, 667–674.

Öller, L.-E., Barot, B., 2000. Comparing the accuracy of European GDP forecasts. International Journal of Forecasting 16.

Ormerod, P., Campbell, M., 1997. Predictability and economic time series, with discussion. In: Haij, C. et al. (Eds.), System Dynamics in Economic and Financial Models. Wiley, Chichester, UK, pp. 73–103.

Pain, N., 1994. Cointegration and forecast evaluation: Some lessons from National Institute forecasts. Journal of Forecasting 13, 481–494.

Pons, J., 2000. The accuracy of IMF and OECD forecasts for G7 countries. Journal of Forecasting 19, 53–63.

Qi, M., 2001. Predicting US recessions with leading indicators via neural network models. International Journal of Forecasting 17, 383–401.

Rich, R.W., 1989. Testing the rationality of inflation forecasts from survey data: Another look at the SRC expected price change data. Review of Economics and Statistics, 682–686.

Samuelson, P.A., 1976. Optimality of sluggish predictors under ergodic probabilities. International Economic Review 17, 1–17.

Schnader, M.H., Stekler, H.O., 1990. Evaluating predictions of change. Journal of Business 63, 99–107.

Schnader, M.H., Stekler, H.O., 1998. Sources of turning point forecast errors. Applied Economics Letters 5, 519–521.

Smith, R., 1994. The macromodelling industry: Structure, conduct and performance. In: Hall, S. (Ed.), Applied Economic Forecasting. Harvester Wheatsheaf, London, pp. 68–88.

Smyth, D., Ash, J.C.K., 1981. The underestimation of forecasts and the variability of predictions and outcomes. Bulletin of Economic Research 33, 37–44.

Spencer, R.W., Huston, J.H., 1993. Rational forecasts—on confirming ambiguity as the mother of conformity. Journal of Economic Psychology 14, 697–709.

Stekler, H.O., 1972. An analysis of turning point errors. American Economic Review 62, 724–729.

Stekler, H.O., 1975. Why do forecasters underestimate? Economic Enquiry 13, 445–449.

Stekler, H.O., 1987. Who forecasts better? Journal of Business and Economic Statistics 5, 155–158.

Stekler, H.O., 1994. Are economic forecasts valuable? Journal of Forecasting 13, 493–505.

Stock, J.H., Watson, M.W., 1999. Forecasting inflation. Journal of Monetary Economics 44, 293–335.

Swanson, N.R., White, H., 1997. Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models. International Journal of Forecasting 13, 439–461.

Tay, A.S., Wallis, K.F., 2000. Density forecasting: A survey. Journal of Forecasting 19, 235–254.

Theil, H., 1966. Applied economic forecasting. In: Economic Forecasting. North-Holland, Amsterdam.

Tkacz, G., 2001. Neural network forecasting of Canadian GDP growth. International Journal of Forecasting 17, 57–69.

Todd, R.M., 1992. Algorithms for explaining forecast revisions. Journal of Forecasting 11, 675–685.

Treasury and Civil Service Committee, 1991. Memoranda on official economic forecasting, 532-i, 11, 675–685.

Turner, D.S., 1990. The role of judgement in macroeconomic forecasting. Journal of Forecasting 9, 315–345.

Wallis, K.F. (Ed.), 1986. Models of the UK Economy 3. Oxford University Press, Oxford.

Wallis, K.F. (Ed.), 1987. Models of the UK Economy 4. Oxford University Press, Oxford.

Wallis, K.F., 1989. Macroeconomic forecasting: A survey. Economic Journal 99, 28–61.

Wallis, K.F., 1993. Comparing macroeconometric models: A review article. Economica 60, 225–237.

Wallis, K.F., Whitley, J.D., 1991. Sources of error in forecasts and expectations: UK economic models, 1984–1988. Journal of Forecasting 10, 231–253.

Webb, R.H., 1995. Forecasts of inflation from VAR models. Journal of Forecasting 14, 267–285.

Young, R.M., 1982. Forecasting with an econometric model: The issue of judgmental adjustment. Journal of Forecasting 1, 189–204.

Zarnowitz, V., 1967. An appraisal of short-term economic forecasts. National Bureau of Economic Research, New York.

Zarnowitz, V., 1979. An analysis of annual and multiperiod quarterly forecasts of aggregate income, output, and the price level. Journal of Business 52.

Zarnowitz, V., 1984. The accuracy of individual and group forecasts from business outlook surveys. Journal of Forecasting 3, 11–26.

Zarnowitz, V., 1985. Rational expectations and macroeconomic forecasts. Journal of Business and Economic Statistics 3, 293–311.

Zarnowitz, V., 1992. Business cycles: Theory, history, indicators and forecasting. In: National Bureau of Economic Research Studies in Business Cycles, vol. 27. University of Chicago Press, Chicago.

Zarnowitz, V., Braun, P., 1992. Twenty two years of the NBER-ASA Quarterly Outlook Surveys: Aspects and comparisons of forecasting performance. NBER Working Paper 3965, New York.

Zellner, A., Chen, B., 2001. Bayesian modeling of economies and data requirements. Macroeconomic Dynamics 5, forthcoming.

Zellner, A., Hong, C., 1989. Forecasting international growth rates using Bayesian shrinkage and other procedures. Journal of Econometrics 40, 183–202.