Forecast Evaluation with Cross-Sectional Data: The Blue Chip Surveys

ANDY BAUER, ROBERT A. EISENBEIS, DANIEL F. WAGGONER, AND TAO ZHA

All of the authors work in the Atlanta Fed's research department. Bauer is an analyst, Eisenbeis is senior vice president and director of research, Waggoner is an economist and assistant policy adviser, and Zha is an assistant vice president and policy adviser.

The authors thank Bevin Janci for her valuable research assistance.

Evaluating the accuracy of economic forecasts is critical if they are to be used in decision making. When a single variable is being forecast by a model with several independent variables, accuracy is typically evaluated using mean square error, mean absolute error, or some similar criterion. When a set of variables is being projected in a simultaneous equation setting, researchers still typically assess forecast accuracy for each dependent variable in the system separately. Though this practice is acceptable as a first pass, it ignores three important aspects of the accuracy assessment process. First, using univariate comparisons to forecast a joint system fails to consider possible correlations in the forecast errors, which might bias the assessment. Second, univariate approaches may not be able to rank forecasters uniquely in terms of their overall performance because one forecaster may perform better on one variable while others may perform better on other variables. This consideration is important because these forecasters are projecting a set of economic variables, which should be internally consistent. Being off on several key dimensions but right on one variable provides some indications about the overall quality of the forecast. Finally, while currently employed statistical comparisons reveal how well models or forecasters may perform on average, they do not help to evaluate and compare particular point forecasts at given times.

Using the methodology developed in Eisenbeis, Waggoner, and Zha (2002), which addresses each of the problems mentioned previously, this article explores and compares the economic forecasts in the Blue Chip Economic Indicators Survey. These data are particularly well suited for the problem at hand. The survey has been published monthly since 1977 and contains forecasts of many macroeconomic variables over a relatively long time span. Although variables have been added or dropped, a substantial number have been present since the survey's inception. The forecasters are a mix of economists from major investment banks, corporations, consulting firms, and academic institutions. On average, the survey contains fifty forecasts each month, and many of the forecasters have participated in the survey for several years. The survey thus provides a useful set of forecasts to explore the methodologies and to investigate several aspects of forecast performance over time.

The article also examines whether several key assumptions underlying the measures advocated in Eisenbeis, Waggoner, and Zha (2002) hold; the results show that these assumptions are satisfied for the Blue Chip data set, at least for longer horizon forecasts. The analysis shows that the Blue Chip Consensus Forecast, which is the average of the individual forecasts, performs better than any individual forecaster although several forecasters performed almost as well as the consensus. This

finding indicates that averaging the forecasts across many forecasters removes some of the noise in each individual forecast. This finding also has implications for combining forecasts from different econometric models, a practice that has been extensively explored in the literature (Bates and Granger 1969; Newbold and Granger 1974; Clemen 1989; de Menezes, Bunn, and Taylor 2000).

The discussion first outlines the methodology used in Eisenbeis, Waggoner, and Zha (2002) and details the Blue Chip data and the benchmark data used to evaluate the forecasts. The article then describes the empirical results and provides some conclusions.

The rank and the score of a forecast are similar in the sense that over time the two measures will be uniformly distributed over some interval.

Methodology

There are many different ways to assess the accuracy of forecasts. Ultimately, determining which forecast is best depends on the use to which it will be put. If accuracy in forecasting output is more important than accuracy in forecasting inflation, then one will want to use forecasts that deliver accurate measures of output relative to inflation. The purpose of this article is to evaluate and compare the general accuracy of a set of multivariate forecasts over time. The methodology in its basic form penalizes errors on easy-to-forecast dimensions more than errors on hard-to-forecast dimensions and considers correlations among the forecast errors.[1] Following Eisenbeis, Waggoner, and Zha (2002), this study uses a composite score based on the standard theory of probability and statistics. This score can be used to compare forecasts even if the number of variables being forecast, or their definitions, changes over time. Finally, the method has the advantage of reducing forecast performance assessment to a single number with an easy interpretation.

In one dimension, the squared error is a standard choice to evaluate and compare forecasts. For example, if y is gross domestic product (GDP) growth and $\hat{y}$ is a forecast of y, then $(y - \hat{y})$ is the forecast error and $(y - \hat{y})^2$ is the squared error. If the forecast error is normal with mean zero and variance σ^2, then the normalized squared error,

$$(1)\quad (y - \hat{y})^2/\sigma^2,$$

has a chi-square distribution with one degree of freedom.[2] With the aid of a chi-square table, one could look up the probability of observing a normalized squared error even larger than the given one. This theoretical probability, converted to a percentage, is the score of a forecast as defined in Eisenbeis, Waggoner, and Zha (2002). Under the assumption of normality, over time the forecast scores would vary uniformly between 0 and 100.[3] This assumption does not mean that the scores of each forecaster would vary uniformly between 0 and 100. A superior forecaster might have scores concentrated in the upper end of this range while an inferior forecaster's scores might lie mostly in the lower end.

To evaluate multivariate forecasts, one simply uses the multivariate generalization of the univariate normal distribution. (See the box on page 20 for a discussion of the multivariate normal distribution.) Suppose that $\mathbf{y}$ is a vector of economic variables to be forecast. For the sake of illustration, suppose that $\mathbf{y}$ consists of only two variables: GDP and the consumer price index (CPI). If $\hat{\mathbf{y}}$ is a forecast vector of the two variables, then the forecast error is the vector of the difference between forecast and realized GDP and forecast and realized CPI, denoted by $(\mathbf{y} - \hat{\mathbf{y}})$. If the forecast error has a multivariate normal distribution with mean 0 and variance $\mathbf{\Omega}$, then the analog of (1) is

$$(2)\quad (\mathbf{y} - \hat{\mathbf{y}})'\Omega^{-1}(\mathbf{y} - \hat{\mathbf{y}}),$$

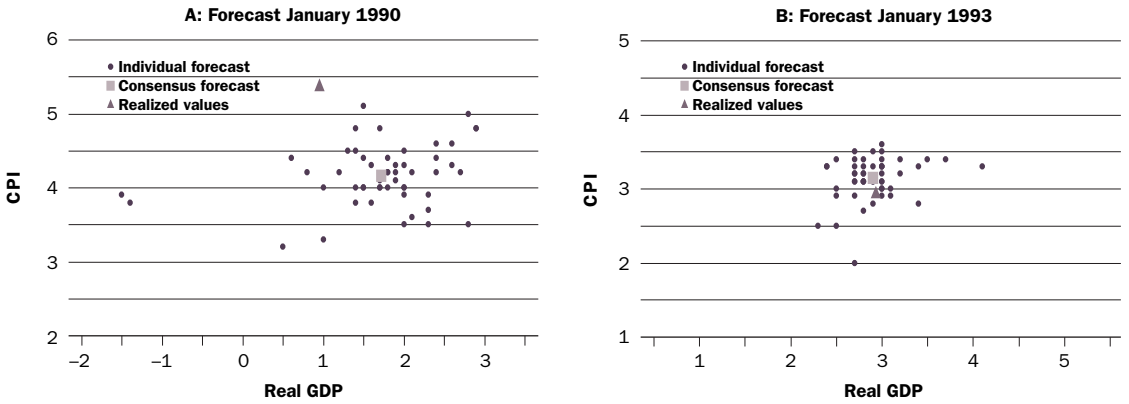
which has a chi-square distribution with degrees of freedom equal to the number of variables.[4] The score of a forecast is defined as the probability, in percentage terms, of observing a normalized squared error even larger than the given one.

The normalized squared error given by equation (1) or (2) is a special case of the more general notion of a loss function. A loss function is simply a mapping of the forecast errors to non-negative numbers and is interpreted as the loss, economic or otherwise, resulting from making a particular error. If the distribution of the loss function applied to the forecast errors were known, then the score of a forecast could be defined as the probability of observing a loss even greater than the loss associated with the given forecast error. This approach is used in Eisenbeis, Waggoner, and Zha (2002).

Loss functions can also be used to rank forecasts. The rank and the score of a forecast are similar in the sense that over time the two measures will be uniformly distributed over some interval; if a forecaster

FIGURE 1

Forecasts of Real GDP and CPI



has superior (or inferior) skill, then the measures for that forecaster will be skewed toward one end of the interval. The difference between these measures is that the rank is always relative to the other forecasters in the group while the score is in absolute terms. If the realized value of the forecast variables is far from the average forecast, then most of the scores will be low while the ranks will always be distributed between 1 and the number of forecasters. Both measures are useful, and both will be reported in this article.

In some contexts there is an obvious candidate for the loss function, but in general there is often no canonical choice. The loss functions given by equations (1) and (2) are called quadratic loss functions and have often been used in the forecasting literature. In univariate models, the choice of σ will have no effect on the forecasts' ranking. However, in multivariate models different choices of Ω will induce different rankings. The matrix Ω determines, among other things, the relative importance of the forecast errors of the individual variables. Assigning different weights to these forecast errors could produce different ranks. This analysis uses assumptions about the distributions of the forecast errors to inform the choice of Ω. Errors in forecasting easy-to-forecast variables are penalized more than errors in forecasting hard-to-forecast variables.

The forecast error variance can be divided into two segments—error variance attributed to unpredictable events and error variance caused by using imperfect forecasting models, also known as model uncertainty. Even if a forecaster had access to all available information at a given time and had perfect foresight in combining this information to make a forecast, the forecast usually would not be equal to the observed values. Unpredictable events occurring after the forecast has been made ensure that no forecast of economic variables can always be exact. The variance of this hypothetical best forecast relative to the realized value is denoted by Ω^H.[5] The consensus forecast can be used to approximate the hypothetical best forecast.[6]

In practice, a forecaster does not have access to all available information or perfect foresight in using information to make forecasts. Thus, an actual forecast will vary from this hypothetical best forecast. This variance is denoted by Ω^F. Figure 1 compares the joint forecasts of GDP and the CPI for two arbitrarily chosen periods.

1. The methodology could easily be generalized to consider the costs of different types of errors.
2. Information about and tables for the chi-square distribution can be found in any elementary statistics text. It is important to note that this definition depends heavily on the assumption of normality of the forecast error. In practice, the forecast error is not exactly normal but is close enough so that this is not an extreme assumption.
3. If x is a normalized squared error and s is its associated score, then the probability of observing a score less than s is equal to the probability of observing a normalized squared error greater than x, which, from the definition of the score, is $s/100$. Thus, the probability of observing a score in any subinterval of (0, 100) is proportional to the length of the subinterval. This proportionality is the defining feature of the uniform distribution.
4. Here, $(\mathbf{y} - \mathbf{y})'$ is a row vector, Ω^{-1} is the matrix inverse of Ω, and $(\mathbf{y} - \hat{\mathbf{y}})$ is a column vector. The product $(\mathbf{y} - \hat{\mathbf{y}})'\Omega^{-1}(\mathbf{y} - \hat{\mathbf{y}})$ is matrix multiplication and results in a single number.
5. In this case, the hypothetical best forecast is also known as the conditional mean of the variables being forecast. The conditional mean minimizes the expected score.
6. Using the language of Blue Chip, the consensus forecast is considered to be the mean, or average, forecast.

BOX

Multivariate Normal and Chi-Square Distributions

A univariate normal distribution is characterized by two numbers, the mean and the variance. The mean centers the distribution, and the variance determines the dispersion. A multivariate normal distribution is also characterized by its mean and variance, but the mean is a vector and the variance is a matrix. The figure shows a sample of 200 points from a two-dimensional normal distribution. The mean of this distribution is (1, 2), and the variance is

$$\begin{pmatrix} 0.81 & 0.27 \\ 0.27 & 0.36 \end{pmatrix}.$$

Each coordinate of a multivariate normal distribution will have a univariate normal distribution. In this case, the mean of the first coordinate is 1 with a variance of 0.81, and the mean of the second coordinate is 2 with a variance of 0.36. The covariance of the two coordinates, which is the correlation times the square root of the variances, is 0.27. Thus, the correlation is $0.27/\sqrt{0.81 \times 0.36} = 0.5$. In general, the elements of a variance matrix are the variance and covariance among the individual coordinates.

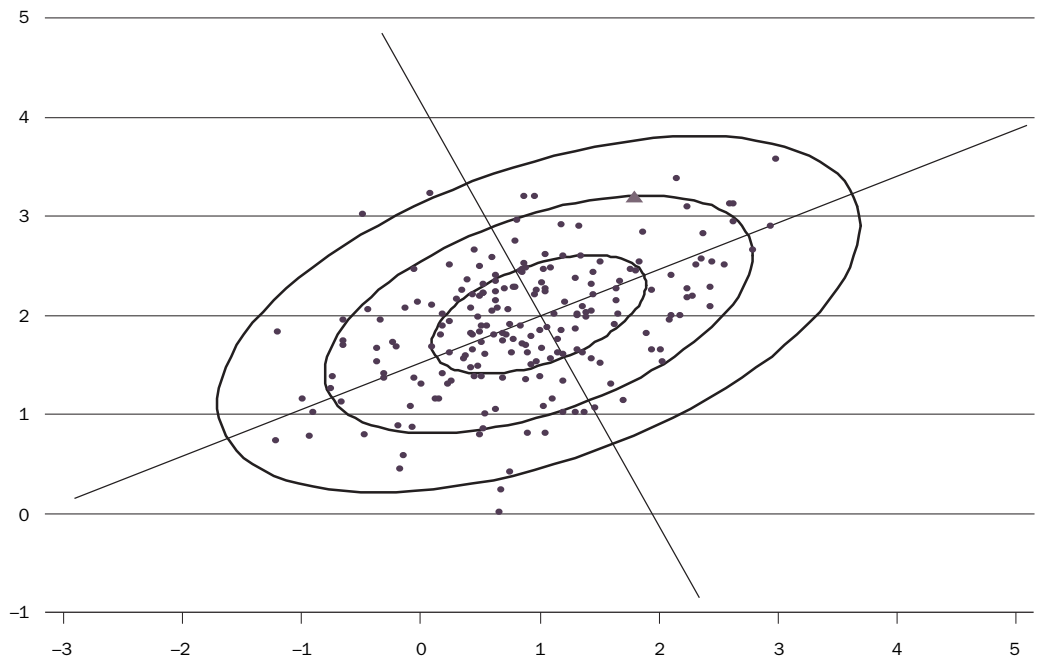
The triangular point on the middle ellipse has coordinates (1.8, 3.2), and the matrix product

$$\begin{pmatrix} 1.8-1 & 3.2-2 \end{pmatrix} \begin{pmatrix} 1.65 & -1.23 \\ -1.23 & 3.70 \end{pmatrix} \begin{pmatrix} 1.8-1 \\ 3.2-2 \end{pmatrix}$$

is approximately 4. The row and column vectors are the difference between the triangular point and the mean while the matrix is the inverse of the variance given above. The middle ellipse has the property that the above product will always be 4 for any point along the ellipse. For points inside this ellipse, the product will be less than 4, and for points outside the ellipse the product will be greater than 4. On the outer ellipse the product will be 9, and on the inner ellipse the product will be 1. By applying the above product to any data point, a two-dimensional normal distribution is transformed into a one-dimensional chi-square distribution with two degrees of freedom. In general, the degrees of freedom will be equal to the number of variables. Indeed, the chi-square distribution is defined in this way from a multivariate normal distribution. Using a chi-square table (or the Microsoft Excel function CHIDIST), one can find the probability that a data point will lie outside a given ellipse or, equivalently, the probability that the product computed from the data point will be greater than some given value. This method defines and should be used to interpret the score of a forecast.

FIGURE

Sample from a Two-Dimensional Normal Distribution



In both panels A and B there is considerable variation between the individual forecasts and the consensus. This variation is captured by Ω^F. The difference between the consensus forecast and the realized values is fairly large in panel A but is smaller in panel B. This variation is captured by Ω^H. The total variation over time will be the sum of these two.

Symbolically, if $\mathbf{y}$ denotes the vector of economic variables being forecast, $\hat{\mathbf{y}}$ is an individual forecast, and $\bar{\mathbf{y}}$ is the hypothetical best or consensus forecast, then the forecast error can be partitioned as

$$\mathbf{y} - \hat{\mathbf{y}} = (\mathbf{y} - \bar{\mathbf{y}}) + (\bar{\mathbf{y}} - \hat{\mathbf{y}}).$$

The first term is the error in the hypothetical best or consensus forecast, and the second term is the additional error due to the difference between the actual forecast and the best forecast. If these two errors are independent, then the variance of the forecast error will simply be the sum of the two corresponding variances, Ω^H and Ω^F. Thus,

$$\Omega = \Omega^H + \Omega^F.$$

Given the large cross section of data in the Blue Chip Survey, Ω^F can easily be estimated as the variance matrix of the forecasts under the assumption that the Blue Chip Consensus Forecast is an acceptable proxy for the hypothetical best forecast (see recent work by Ottaviani and Sorensen 2003). The matrix Ω^H can be estimated as the variance matrix of the realized forecast errors of the consensus forecast. About fifty forecasts are in each Blue Chip Survey, enough to estimate Ω^F. However, only one Blue Chip Consensus Forecast exists for each period, so a relatively long series of forecasts is needed to estimate Ω^H. There must be at least as many consensus forecasts as there are variables being forecast, and ideally there should be more than three to four times as many consensus forecasts as variables. In Eisenbeis, Waggoner, and Zha (2002), a particular forecasting model was estimated, and Ω^H was taken to be the theoretical variance matrix from this model. This estimation was necessary because the variables used in the *Wall Street Journal* forecasts frequently changed over time, and so there was not a long enough time series of mean forecasts. Both methods for obtaining Ω^H are considered here to compare the sensitivity of the proposed performance measures to differences in estimates of Ω^H.

The Data

The Blue Chip Economic Indicators Survey has been published monthly for more than twenty-five years. The survey includes the annual average of, or change in, fifteen macroeconomic variables. Forecasts of five variables are considered here: real GDP, the CPI, the unemployment rate, three-month Treasury bill rates, and ten-year Treasury note yields. With the exception of the ten-year Treasury note, these variables have been included in all of the surveys. Prior to 1996, a corporate bond yield was forecast instead of the Treasury note. Though differences exist between Treasury and corporate yields, these series are joined for illustrative purposes in this study. Approximately fifty firms participate in each survey. Though the number of firms has remained roughly constant, the identities of the firms have changed over time. In some instances, firms merged or ceased to appear in the survey for various reasons. The analysis tracks merged firms and combined forecasts to create long time series when possible. The participation dates of each firm and mergers are noted in Table 1.

The forecast error variance can be divided into two segments—error variance attributed to unpredictable events and error variance caused by using imperfect forecasting models, also known as model uncertainty.

Understanding the dating conventions of the forecasts is important for understanding the tables, figures, and discussions in this article. Near the beginning of every month each forecaster submits two forecasts: one for the current calendar year and another for the next calendar year. For instance, in January 2000, forecasters submit a forecast for 2000 (current year) and for 2001 (next year). Forecasts are dated by the year and month in which they are made and identified as either current or next year. Next-year forecasts made in January are long-term forecasts. These forecasts will not be completely realized for twenty-four months. On the other hand, current-year forecasts made in December are short-term forecasts that will be realized in one month. For each year being forecast, twenty-four forecasts with horizons varying from one to twenty-four months are made. This study includes current-year forecasts from January 1986 through December 2001 and next-year forecasts from January 1986 through December 2000.

To determine the accuracy of the forecasts, benchmark or realized values of the variables must

TABLE 1

Average Scores

	Years in Survey[1]	Average Score		Current-Year Average Score		Next-Year Average Score	
BC Consensus	86–01	69.3***	(21.8)	74.7***	(23.2)	63.5**	(18.7)
Security Pacific National Bank	86–92	68.8**	(24.4)	73.9**	(27.2)	63.6	(19.8)
NationsBank	93–98	67.7*	(23.0)	67.7*	(27.7)	67.7*	(17.3)
Mortgage Banker Assn. of America	86–01	67.1***	(25.9)	73.2***	(26.4)	60.7*	(23.8)
Macroeconomic Advisors[2]	86–01	66.6**	(25.9)	74.2***	(25.4)	58.4	(24.0)
U.S. Trust Company	86–01	63.6**	(26.0)	68.3***	(25.7)	56.1	(24.8)
CoreStates Financial Corporation	88–98	63.0*	(24.4)	61.4*	(28.2)	64.7**	(19.7)
Pennzoil Company	86–89, 92–93	62.9	(26.4)	68.0*	(28.0)	57.7	(23.9)
Northern Trust Company	86–01	62.7**	(26.6)	66.0**	(26.9)	58.9	(25.7)
Bank of America	87–01	62.7**	(26.2)	64.7**	(28.5)	60.6*	(23.5)
Equitable Life Assurance	86–91	62.5	(26.3)	69.6**	(25.0)	52.4	(25.0)
Peter L. Bernstein, Inc.	86–89	62.2	(28.8)	64.8	(28.5)	59.6	(29.3)
Moody's Investors Service	98–01	61.7	(26.7)	66.5	(31.2)	55.0	(17.2)
Wayne Hummer Investments, LLC	86–01	60.6*	(25.7)	62.5**	(27.5)	58.6	(23.6)
Merrill Lynch	86–01	60.1*	(26.3)	64.2**	(27.2)	55.5	(24.5)
Dean Witter Reynolds & Company	86–91	60.0	(30.0)	63.0	(30.7)	56.1	(29.1)
PNC Financial Corporation	88–98	59.3	(24.3)	59.2	(27.6)	59.3	(20.3)
Fleet Financial Group[3]	91–99	58.8	(24.1)	62.7*	(28.5)	54.9	(18.0)
Metropolitan Life Insurance Company	86–96	58.7	(26.0)	59.7	(31.8)	57.7	(18.6)
Wells Capital Management[4]	91–01	58.2	(27.2)	62.8*	(29.0)	53.1	(24.3)
Georgia State University	86–01	58.2	(26.3)	59.2	(28.1)	57.2	(24.3)
National Association of Home Builders	90–01	58.1	(24.3)	62.8*	(25.7)	53.0	(21.5)
Chicago Capital, Inc.	96–00	57.6	(32.5)	63.3	(31.4)	51.7	(32.8)
University of Michigan M.Q.E.M.	86–96	56.9	(28.7)	67.5**	(28.1)	46.3	(25.3)
National City Corporation[5]	86–01	56.8	(24.0)	59.4*	(25.6)	54.0	(21.9)
Evans Group	86–01	56.5	(28.6)	63.4**	(28.7)	49.1	(26.7)
Eggert Economic Enterprises, Inc.	86–01	56.5	(24.3)	55.6	(27.1)	57.5	(21.1)
DaimlerChrysler AG[6]	86–01	56.3	(28.1)	62.8**	(28.1)	49.4	(26.3)
Chase Manhattan Bank	88–00	56.2	(28.7)	61.3*	(29.1)	50.3	(27.2)
La Salle National Bank	86–91, 97–01	56.1	(28.0)	62.0*	(30.1)	49.4	(24.0)
Dun & Bradstreet	89–99	55.9	(28.0)	59.4	(29.0)	52.3	(26.5)
DuPont	86–01	55.7	(26.0)	59.3*	(28.8)	51.9	(22.0)
Bank One[7]	86–01	55.3	(30.7)	61.1*	(31.1)	48.9	(29.0)
Siff, Oakley, Marks, Inc.	86–01	54.8	(27.5)	60.9*	(26.4)	48.2	(27.2)
Charles Reeder	86–99	54.6	(29.0)	54.7	(32.1)	54.4	(25.6)
Bear Stearns & Company, Inc.	97–01	54.0	(31.9)	54.5	(32.9)	53.1	(30.7)
Standard & Poor's	94–01	54.0	(28.4)	61.5	(30.4)	45.5	(23.2)
Prudential Financial[8]	86–01	54.0	(25.7)	55.7	(28.1)	52.1	(22.7)
Fannie Mae	98–01	53.3	(26.5)	59.7	(28.5)	44.8	(21.2)
U.S. Chamber of Commerce	86–01	51.9	(26.7)	54.8	(27.5)	48.7	(25.6)
Sears, Roebuck and Company	86–95	51.7	(28.3)	56.8	(30.3)	46.5	(25.1)
Motorola	96–01	51.0	(28.6)	58.5	(31.1)	42.2	(22.7)
UCLA Business Forecast	86–01	49.9	(28.8)	50.9	(31.5)	48.8	(25.8)
Wachovia Securities[9]	96–01	49.8	(25.0)	53.9	(27.5)	44.7	(20.5)
General Motors Corporation	92–01	49.1	(26.0)	47.1	(29.8)	51.3	(21.1)
Comerica[10]	90–01	48.9	(25.6)	49.0	(30.4)	48.8	(19.3)
Goldman Sachs & Company	98–01	47.8	(29.4)	63.8	(25.9)	25.5**	(16.9)
Econoclast	86–01	47.5	(27.0)	45.6	(31.0)	49.5	(21.9)
Prudential Securities[11]	86–96, 00–01	46.6	(31.6)	46.8	(34.0)	46.2	(27.6)
Conference Board	86–01	46.4	(29.7)	53.8	(32.1)	38.4*	(24.5)
Turning Points (Micrometrics)	89–01	46.0	(27.8)	44.2	(30.6)	47.8	(24.4)
Eaton	94–01	45.4	(27.6)	40.9	(28.8)	50.5	(25.4)
JPMorgan Chase[12]	96–01	44.3	(27.9)	52.8	(29.4)	34.3*	(22.3)

	Years in Survey[1]	Average Score		Current-Year Average Score		Next-Year Average Score	
Cahners Publishing Company	86–98	43.0	(25.2)	47.2	(27.8)	38.7*	(21.6)
DRI-WEFA[13]	98–01	42.2	(28.8)	51.7	(29.3)	29.0*	(22.3)
Fairmodel Economica, Inc.	86–93	41.9	(31.5)	45.9	(33.6)	37.9	(28.9)
Chemical Banking[14]	86–95	41.6	(28.9)	45.4	(29.1)	37.1*	(28.1)
Kellner Economic Advisers	97–01	40.9	(20.6)	44.0	(23.7)	37.0	(15.1)
Weyerhaeuser Company	94–00	40.3	(25.1)	42.7	(29.1)	37.8	(20.2)
C.J. Lawrence, Inc.	91–96	39.7	(28.6)	27.9**	(26.0)	56.3	(23.5)
Polyconomics	86–89	38.7	(27.2)	39.7	(29.2)	37.7	(25.4)
Genetski Financial Advisors	92–95, 01	38.3	(30.4)	53.1	(31.3)	21.4**	(18.1)
Morris Cohen & Associates	86–96	37.4*	(28.9)	22.6***	(24.1)	53.7	(24.8)
Bostian Economic Research	86–97	37.0*	(28.6)	26.1***	(28.5)	47.9	(24.2)
Arnhold & S. Bleichroeder	86–93	36.8*	(32.8)	28.5**	(29.5)	46.4	(34.0)
Ford Motor Company	96–01	36.7	(27.7)	37.8	(28.9)	35.0	(26.2)
Inforum–University of Maryland	86–01	36.6**	(26.6)	33.6**	(27.1)	39.8*	(25.8)
Deutsche Banc Alex. Brown[15]	96–01	36.5	(33.1)	37.8	(30.5)	34.6*	(37.0)
Econoviews International, Inc.	86–92	36.3	(28.5)	35.0*	(30.4)	37.7	(26.7)
Morgan Stanley	97–01	33.5	(27.1)	34.4	(29.9)	31.3*	(19.2)
Business Economics, Inc.	86–89	14.3***	(16.5)	13.7***	(16.6)	14.9***	(16.4)

Note: The table shows the average score of forecasters with at least four years of data—seventy forecasters out of a total sample of one hundred four. Numbers in parentheses are standard deviations. *, **, and *** represent significance at the 90 percent, 95 percent, and 99 percent confidence levels, respectively.

1. Years in which there were at least four monthly forecasts for the five variables evaluated in this article.
2. Prior to 07/96, forecasts were from Meyer & Associates.
3. Prior to 12/95, forecasts were from Shawmut National Corporation.
4. Prior to 09/01, forecasts were from Wells Fargo and before 06/96 were from First Interstate Bancorp.
5. Prior to 01/00, forecasts were from National City Bank of Cleveland.
6. Prior to 09/01, forecasts were from Chrysler Corporation.
7. Prior to 11/98, forecasts were from First National Bank of Chicago.
8. Prior to 08/01, forecasts were from Prudential Insurance.
9. Prior to 11/01, forecasts were from First Union Corporation.
10. Prior to 08/92, forecasts were from Manufacturers National Bank of Detroit.
11. Prior to 01/92, forecasts were from Prudential Bache Securities.
12. Prior to 09/01, forecasts were from JPMorgan.
13. Prior to 09/01, forecasts were from WEFA Group.
14. Prior to 02/92, forecasts were from Manufacturers Hanover Trust.
15. Prior to 09/01, forecasts were from Deutsche Morgan Grenfell.

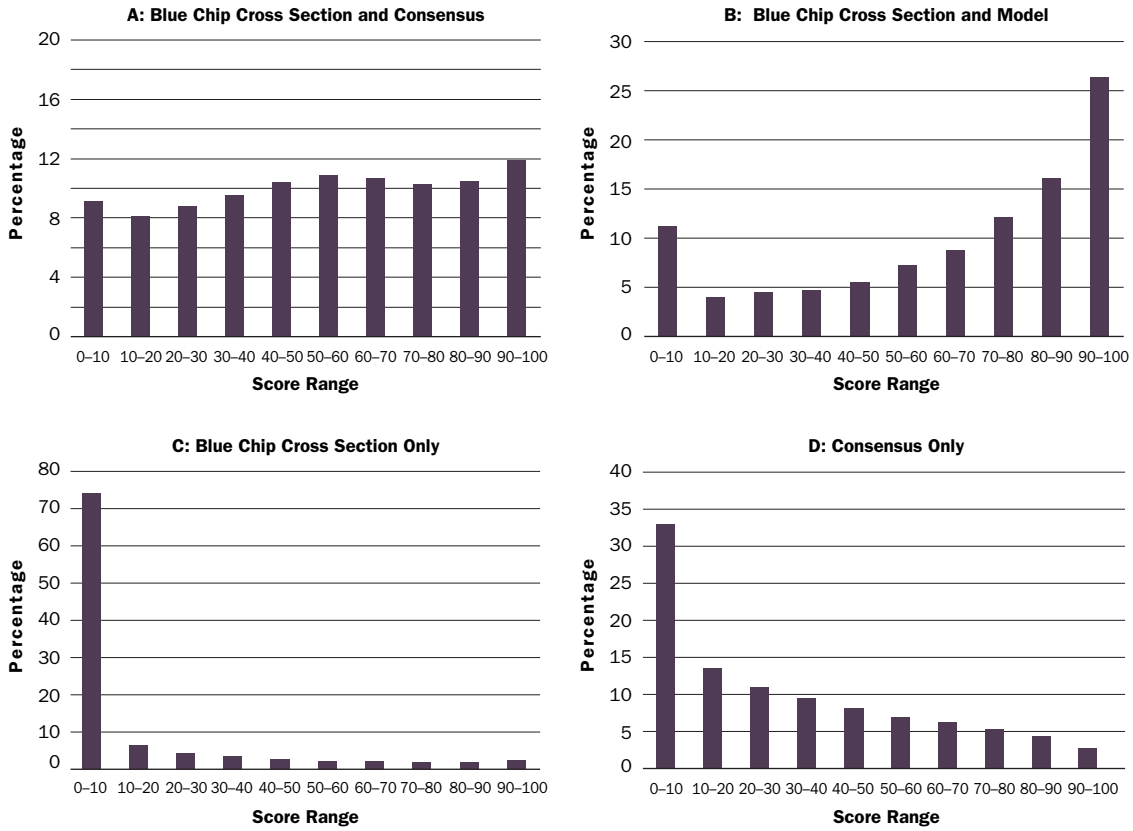
be available. The appropriate choice of a benchmark is complicated by the fact that some series are revised over time. For example, GDP is reported quarterly and revised twice. The advance number is reported in the first month following the end of the quarter, the revised preliminary number is released the next month, and the final number appears three months after the end of the quarter. Also, every July additional revisions may be made to past data. In addition, changes in the definitions of these series may be made. For example, in January 1996 the Bureau of Economic Analysis changed measurement of GDP to a chain-weighted system. This change could be responsible for some of the poor forecasting results observed at the end of 1995 because the forecasts made in 1994 and 1995 for GDP growth over 1995 would be based on the non-chain-weighted series while the GDP data available to assess them would use the chain-weighted numbers. These issues make it important to use vintage data when accessing the accuracy of past forecasts. Vintage, or real-time, data are the data available to the forecaster at a specific time. For instance, vintage January 1990 data are the data that were available to a forecaster at the end of January 1990. For a revised series such as GDP, vintage data would be the advance number for the last quarter of 1989 and the final number for previous quarters. The series used to evaluate forecast accuracy are described in detail in the appendix.

Variance Estimates

As the discussion of methodology showed, if the distribution of the forecast errors is approximately normal, then over time the scores would be

FIGURE 2

Blue Chip Forecast Scores



approximately uniformly distributed. Conversely, an approximately uniform distribution of the scores would be evidence that the underlying assumptions were not grossly violated. This uniformity is important if one is to take seriously the interpretation of the score as the percentage of forecasts expected to be worse than the given one. The uniformity of the distribution of scores will also be sensitive to the choice of estimate of Ω.

Figure 2 plots a histogram of the scores for both the current and next year. Four different variance matrices are used. In panel A the variance is the sum of the cross-sectional variance of the Blue Chip Survey and the variance of the forecast error of the Blue Chip Consensus Forecast. This is the baseline case. In panel B, the variance is the sum of the cross-sectional variance of the Blue Chip Survey and the estimate of the forecast error variance from the theoretical model used in Eisenbeis, Waggoner, and Zha (2002). In panel C only the cross-sectional variance of the Blue Chip Survey is used, and in panel D only the variance of the forecast error of the Blue Chip Consensus is used.

The plot in panel A is virtually uniform, as desired. The histogram in panel B is less uniform and skewed toward higher scores, indicating that the forecast error variance from the model may be too large.[7] Both panels C and D are highly skewed toward lower scores, with the histogram associated with the cross-sectional variance of the Blue Chip Survey in panel C the more skewed. This skewness indicates that individually neither of these matrices captures all of the forecast error variance and that the cross-sectional variance is smaller than the variance of the consensus forecast error. This result is consistent with that of Zarnowitz and Lambros (1987), which showed that in the univariate case the cross-sectional variance underestimates the overall uncertainty of forecasts. It is often claimed that professional forecasters are looking over each other's shoulders and thus produce similar forecasts. The results here are consistent with this view, but they are also consistent with the view that forecasters are all making forecasts close to some hypothetical best forecast but that this best forecast may not be that close to the realized values.

TABLE 2

Score Distribution by Forecast Horizon

	Count	Percentage of Forecast Scores in Each Range									
		0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80	80–90	90–100
					Current Year						
December	752	16.1	4.8	2.7	3.9	5.1	6.6	4.1	7.6	10.2	39.0
November	750	16.8	4.5	3.5	4.3	4.8	7.1	10.7	9.1	9.5	29.9
October	754	15.5	5.0	4.5	4.0	5.4	5.2	6.4	9.5	17.1	27.3
September	752	13.8	6.4	6.6	5.5	7.0	6.5	6.4	10.1	13.8	23.8
August	745	12.2	7.0	7.5	5.8	9.3	7.7	7.5	9.4	13.4	20.3
July	760	11.1	9.3	7.4	8.7	8.3	10.3	9.3	8.7	12.2	14.7
June	749	9.3	8.3	8.4	8.1	7.7	8.7	10.7	13.5	10.5	14.7
May	743	9.3	7.0	8.9	9.4	9.4	9.0	13.2	11.3	11.0	11.4
April	754	7.6	8.8	8.6	9.2	10.9	11.0	12.3	11.4	12.1	8.2
March	752	8.0	8.2	10.9	10.5	11.2	13.7	10.8	10.9	8.9	6.9
February	739	7.2	9.2	10.1	14.6	10.7	13.3	10.6	8.7	9.2	6.5
January	734	9.5	7.2	8.2	9.5	13.1	12.5	13.6	10.9	9.4	6.0
					Next Year						
December	703	8.7	9.5	7.4	10.1	13.8	12.2	11.0	12.1	9.7	5.5
November	698	7.7	8.7	8.0	9.9	11.5	14.2	12.9	12.6	8.9	5.6
October	705	7.1	8.7	8.1	11.9	10.8	13.9	12.6	11.1	11.1	4.8
September	698	6.2	10.6	8.5	11.3	11.7	11.6	12.3	10.7	10.9	6.2
August	694	6.8	8.8	9.5	9.1	11.4	13.8	13.4	10.1	11.2	5.9
July	699	6.3	9.3	10.9	10.6	11.0	13.9	11.4	10.0	9.7	6.9
June	687	6.3	8.3	10.6	13.0	12.2	11.2	11.2	10.2	11.5	5.5
May	663	5.6	7.8	13.3	11.5	12.1	11.8	13.4	11.0	8.9	4.7
April	667	5.1	9.1	12.0	12.3	12.9	12.4	12.3	9.6	8.7	5.5
March	641	5.8	8.7	11.7	12.9	13.3	13.4	10.6	9.2	8.6	5.8
February	599	5.8	10.9	11.9	13.9	15.7	10.7	10.4	8.8	6.5	5.5
January	553	7.4	9.6	14.5	11.2	14.5	13.2	9.0	9.8	4.9	6.0
All	16,991	9.1	8.1	8.7	9.5	10.4	10.9	10.6	10.3	10.5	11.9

Note: Count is the total number of forecasts in the sample for each month and forecast year. Under the assumptions in this article, all percentages should be approximately equal to 10.

Table 2 shows the distributions for the twenty-four forecast horizons using the baseline estimate for the variance. The last row of this table gives the percentages for the histogram presented in panel A of Figure 2. The other rows can be interpreted as histograms for each forecast horizon. For longer horizons, the scores are approximately uniformly distributed, but for shorter horizons, August through December of the current year, the distribution appears to be U-shaped. A possible explanation for this distribution is that, as the forecast horizon decreases, firms spend fewer resources on the current-year forecast and more on the next-year forecast. In December clients are probably more interested in accurate forecasts of the next calendar year than they are in forecasts of the year that is almost completed. This explanation would account for the higher-than-expected frequency of scores in the lowest range. In turn this pattern would increase and distort the estimate of the variance, which would improve the scores of firms that have good short-term forecasts and explain the higher-than-expected frequency of scores in the high range. Table 2 indicates that this methodology works better for the longer-horizon Blue Chip forecasts.

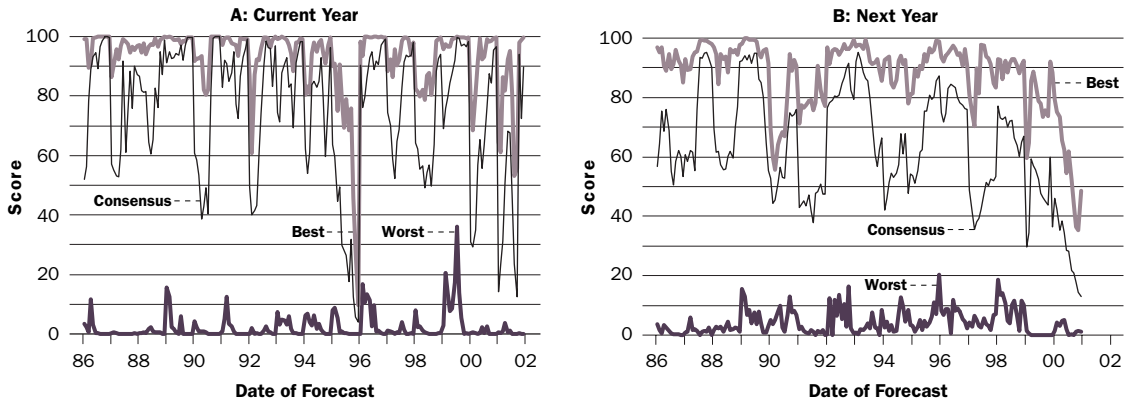
Forecast Performance

Figure 3 plots the score for the consensus forecast for both the current year (panel A) and the

7. In some sense this pattern should be expected since the forecast error variance from the model may be a better proxy for all of Ω instead of just Ω^H. In cases when a too-short time series makes it impossible to use the consensus forecast error variance as a proxy for Ω^H, it may be better to scale the estimate of Ω^H from the model.

FIGURE 3

Blue Chip Consensus Forecast Scores



next year (panel B). The highest and lowest scores for each month are also plotted for comparison. Though the consensus scores vary considerably from month to month, most of the values are above 50 percent. In fact, the average score for the consensus was 75 percent for the current year and 64 percent for the next year. This result means that the consensus forecast was on average more accurate than 75 percent of the current-year forecasts and 64 percent of the next-year forecasts. The three notable exceptions are the current-year forecasts made toward the end of 1995 and the both the current-year and next-year forecasts for 2001 (made in 2001 and 2000, respectively). The low current-year scores toward the end of 1995 are mostly a result of errors in the forecast of GDP, which, as mentioned previously, may stem from the change to a chain-weighted measurement of GDP in 1996. Because forecasts were based on 1995 numbers but all the numbers necessary to evaluate the 1995 forecasts were not available until January 1996, a bias may have been introduced because forecasters were not certain how to adjust their forecasts for the difference in the GDP measure being forecast. This bias is not evident in longer-term forecasts—perhaps because it was small relative to the longer-term forecast errors.

In the forecasts of 2001 made in 2000 (Figure 3, panel B), the forecast errors were large for all the variables except the CPI, with the largest errors occurring in short-term interest rates and GDP. Unlike the 1995 episode, this result can be characterized simply as most of the forecasters having missed the turning point. The 2001 current-year forecasts are a little more complex but are still an interesting case study. The low scores early in the year were caused mainly by large errors in short-term interest rates and GDP, similar to the forecasts of 2001 made in 2000. Early in the year, forecasts were revised and the scores improved. The unforeseen terrorist attack on September 11 caused the economy to be weaker than expected in the last quarter of 2001. This weakness certainly affected the scores of all forecasts, but the largest effect was in current-year forecasts made during the third quarter of 2001. After September 11, forecasts were again revised and were then relatively accurate. This scenario clearly illustrates how economic shocks can cause large swings in the forecast performance. A shock causes prior forecasts to be more inaccurate than they would otherwise be and results in significant revisions, which improve subsequent forecasts.

Table 1 presents the average score of those forecasters with at least four years of data. This criterion leaves seventy forecasters out of one hundred four forecasters in the total sample. Interestingly, out of these seventy forecasters the Blue Chip Consensus Forecast has the highest average score though the average score of several forecasters is almost as good. This result is consistent with the claim that the consensus forecast is a proxy for the hypothetically best forecast and is an argument for giving more weight to the consensus score than to the forecast of any one forecaster.[8]

To further interpret Table 1, note that if a forecaster has average skill, then the mean of T independent scores will be approximately normal with a mean of 50 and a standard deviation of $100/\sqrt{12T}$.[9] For a firm that has been in the sample the entire sixteen years, the average score could be computed on the basis of as many as 372 observations. These observations are not independent because one month's score will be highly correlated with the next month's score. However, forecasts of different

years should be approximately independent. Thus, using T as the number of years a firm has been in the sample gives a more plausible estimate for the standard deviation of the average score reported in Table 1. At the 95 percent confidence level, scores that are more than 1.7 standard deviations apart can be considered statistically different. Putting all of these factors together reveals that, for firms that have been in the sample for the entire sixteen years, scores that are more than 13 percentage points apart can be considered statistically different. Also interesting is the fact that forty-one of the seventy forecasters have average scores that are better than 50 percent, and some of these forecasters have quite long forecasting histories. These figures show that many forecasters have performed consistently well. Conversely, some of the forecasters have scores well below 50 percent and have consistently underperformed.[10]

Table 3 presents the average rank of those forecasters with at least four years of data. Though the exact number of forecasters changes from month to month, the average number of forecasters is approximately forty-seven, and in almost all months there were between forty-two and fifty-two forecasters. For this reason, it is not necessary to scale the ranks to some common interval before averaging. Again, the consensus forecast has the best average rank although several forecasters are close. However, the standard deviation of the consensus forecast rank is less than half that of the others. This result implies that the consensus is consistently among the best forecasts even when its score is relatively low. Figure 4 illustrates this finding. The consensus forecast rank is plotted with the Macroeconomic Advisors' forecast rank for both the current- and next-year forecasts. These figures show how much more volatile the Macroeconomic Advisors' ranks are as compared to the consensus forecast ranks. The plots for all the other top-ranked forecasters are similar.

Improving the Consensus

Instead of using the consensus forecast, would it be better to form a "superconsensus" using only highly ranked forecasters? Table 4 shows the results from using the best forecasters from recent years. The table groups the results of the best one, three, five, ten, fifteen, and twenty-five forecasters for periods from one to five years. The performance of the superconsensus can be compared to that of the regular consensus. Only data available at the time the superconsensus is formed are used in its construction. Panel A compares the average scores of various superconsensuses, and panel B compares the average ranks. These results imply that there is at best a very small gain in average score or rank in using a superconsensus forecast, but there is an increase in the standard deviation of the rank. More interesting is the observation that if only a few forecasters are used, then it is clearly best to use those with a long track record of superior forecasts. However, if more than five forecasts are averaged, there seems to be little advantage to using more than the prior two years to select the best forecasters.

The results in this article indicate that forecasters are all making forecasts close to some hypothetical best forecast but that this best forecast may not be that close to the realized values.

Conclusions

A consistent evaluation of forecasts over time that also respects their multivariate character is essential if the forecasts are to be used for decision making. Having both a cross section and a time series of forecasts, as in the Blue Chip Survey, gives one the ability to perform such an evaluation. The methodology developed in Eisenbeis, Waggoner, and Zha (2002) gives consistent results for the Blue Chip Survey Forecasts, particularly at longer forecast horizons. Furthermore, the methodology reveals that the Blue Chip Consensus Forecast consistently performs better than any of the individual forecasters do. This result is a "reverse Lake Wobegon" effect: none of the forecasters are better than the average forecaster. While no forecaster had a higher average score than the consensus forecast, several were indistinguishably close, and many had average scores well above 50 percent. There are superior forecasters, but no individual has access to all of the independent information from all of the forecasts that is incorporated into the consensus forecast.

8. The result is also consistent with the Ottaviani and Sorensen (2003) hypothesis that the forecasts are unbiased.
9. The mean of a uniform random variable on (0, 100) is 50, and its standard deviation is $100/\sqrt{12}$. The standard deviation of the mean of T independent random variables, each with standard deviation σ, is $\sigma/\sqrt{T}$.
10. This pattern also suggests that there may be some survivorship effects in the data.

TABLE 3

Average Rank

	Years in Survey	Average Rank		Current-Year Average Rank		Next-Year Average Rank	
BC Consensus	86–01	13.2***	(4.8)	12.4***	(5.1)	13.9***	(4.3)
Moody's Investors Service	98–01	13.5*	(10.3)	16.4	(11.0)	9.4**	(7.5)
Security Pacific National Bank	86–92	14.3**	(10.9)	14.4**	(11.6)	14.3**	(10.3)
NationsBank	93–98	15.1**	(12.9)	16.3*	(12.8)	13.8**	(13.0)
Mortgage Banker Assn. of America	86–01	15.5***	(10.4)	14.3***	(9.9)	16.7**	(10.7)
Macroeconomic Advisors	86–01	15.5***	(10.8)	13.7***	(9.9)	17.3**	(11.5)
Northern Trust Company	86–01	17.9**	(12.0)	18.3**	(11.6)	17.5**	(12.4)
Bank of America	87–01	18.0**	(11.5)	19.3*	(12.4)	16.6**	(10.3)
U.S. Trust Company	86–01	18.5**	(12.7)	17.6**	(12.0)	20.0*	(13.6)
CoreStates Financial Corporation	88–98	19.2*	(12.3)	22.5	(13.1)	15.8**	(10.5)
Peter L. Bernstein, Inc.	86–89	19.6	(12.6)	20.5	(12.8)	18.8	(12.4)
Equitable Life Assurance	86–91	19.9	(11.3)	17.8	(11.7)	23.0	(10.0)
Wayne Hummer Investments, LLC	86–01	20.2*	(11.3)	21.9	(12.3)	18.3**	(9.9)
Chicago Capital, Inc.	96–00	20.5	(15.5)	21.0	(13.8)	20.1	(17.2)
Fannie Mae	98–01	20.9	(10.9)	20.8	(11.2)	21.1	(10.7)
Pennzoil Company	86–89, 92–93	21.0	(10.9)	19.7	(12.1)	22.2	(9.5)
Merrill Lynch	86–01	21.0	(12.4)	21.1	(12.3)	20.9	(12.4)
Dean Witter Reynolds & Company	86–91	21.0	(13.1)	22.1	(11.7)	19.6	(14.6)
Wells Capital Management	91–01	21.3	(12.7)	20.7	(11.3)	22.1	(14.2)
Georgia State University	86–01	21.4	(12.7)	23.1	(13.0)	19.6*	(12.2)
National Association of Home Builders	90–01	21.6	(10.8)	21.9	(11.1)	21.2	(10.5)
PNC Financial Corporation	88–98	21.7	(11.1)	23.7	(11.5)	19.6	(10.3)
DaimlerChrysler AG	86–01	21.9	(12.7)	20.5	(12.3)	23.3	(13.0)
Bear Stearns & Company, Inc.	97–01	22.2	(17.1)	24.3	(16.9)	18.8	(17.3)
La Salle National Bank	86–91, 97–01	22.3	(12.3)	21.6	(12.8)	23.2	(11.7)
National City Corporation	86–01	22.4	(11.8)	23.3	(12.3)	21.4	(11.1)
Fleet Financial Group	91–99	22.4	(11.8)	22.0	(12.8)	22.8	(10.7)
Eggert Economic Enterprises, Inc.	86–01	22.4	(12.1)	25.6	(12.0)	19.1*	(11.4)
Evans Group	86–01	22.4	(13.6)	20.7	(13.2)	24.3	(13.7)
Metropolitan Life Insurance Company	86–96	22.6	(11.0)	23.3	(12.8)	21.9	(8.8)
DuPont	86–01	22.8	(11.8)	23.4	(12.1)	22.2	(11.5)
University of Michigan M.Q.E.M.	86–96	22.8	(12.3)	19.3*	(11.7)	26.2	(12.0)
Standard & Poor's	94–01	22.9	(13.6)	19.9	(13.5)	26.3	(12.8)
Dun & Bradstreet	89–99	23.2	(13.9)	22.7	(14.4)	23.8	(13.5)
Bank One	86–01	23.2	(14.9)	21.4	(14.3)	25.1	(15.4)
Wachovia Securities	96–01	23.3	(14.0)	24.9	(14.7)	21.4	(13.0)
Siff, Oakley, Marks, Inc.	86–01	23.5	(12.8)	23.3	(12.0)	23.7	(13.6)
Chase Manhattan Bank	88–00	23.7	(14.3)	22.8	(13.5)	24.8	(15.2)
Prudential Insurance	86–01	23.8	(12.7)	26.0	(12.5)	21.3	(12.5)
Charles Reeder	86–99	23.8	(14.9)	25.7	(15.3)	21.8	(14.2)
Goldman Sachs & Company	98–01	24.5	(14.6)	16.8	(13.0)	35.2*	(8.8)
U.S. Chamber of Commerce	86–01	25.3	(11.7)	26.5	(11.1)	23.9	(12.3)
Sears, Roebuck and Company	86–95	25.7	(12.4)	24.5	(13.0)	26.9	(11.8)
Motorola	96–01	25.9	(12.2)	24.7	(13.3)	27.2	(10.8)
Comerica	90–01	26.1	(13.4)	28.3	(13.4)	23.7	(13.0)
UCLA Business Forecast	86–01	26.2	(13.7)	28.0	(13.7)	24.4	(13.5)
General Motors Corporation	92–01	27.0	(12.9)	29.3	(13.9)	24.4	(11.2)
Prudential Securities	86–96, 00–01	27.2	(13.7)	26.5	(14.7)	28.3	(11.9)

	Years in Survey	Average Rank		Current-Year Average Rank		Next-Year Average Rank	
Econoclast	86–01	27.7	(12.7)	30.5*	(12.7)	24.7	(12.0)
Turning Points (Micrometrics)	89–01	27.7	(12.9)	30.6*	(12.8)	24.6	(12.3)
Eaton	94–01	27.7	(14.0)	32.8*	(11.9)	22.0	(14.0)
Conference Board	86–01	29.2	(13.3)	27.2	(13.8)	31.3**	(12.5)
Kellner Economic Advisers	97–01	29.3	(11.9)	31.5	(10.7)	26.4	(12.9)
DRI-WEFA	98–01	29.8	(12.9)	27.8	(13.1)	32.5	(12.4)
JPMorgan Chase	96–01	29.9	(13.2)	26.1	(15.3)	34.4*	(8.4)
Fairmodel Economica, Inc.	86–93	30.4	(13.9)	30.2	(14.2)	30.6	(13.7)
C.J. Lawrence, Inc.	91–96	30.5	(14.7)	34.4*	(13.9)	25.0	(14.1)
Arnhold & S. Bleichroeder	86–93	30.8	(15.6)	35.8**	(12.3)	25.1	(17.1)
Cahners Publishing Company	86–98	30.9*	(10.9)	30.2*	(11.4)	31.7**	(10.3)
Polyconomics	86–89	31.0	(13.0)	31.9	(12.8)	30.2	(13.3)
Chemical Banking	86–95	31.1*	(13.1)	29.9	(12.0)	32.5*	(14.2)
Bostian Economic Research	86–97	31.1*	(14.9)	35.6***	(14.6)	26.6	(13.8)
Inforum–University of Maryland	86–01	31.1**	(13.5)	33.9***	(11.9)	28.0	(14.5)
Genetski Financial Advisors	92–95, 01	31.2	(13.9)	24.9	(14.0)	38.4**	(9.7)
Weyerhaeuser Company	94–00	31.8	(12.9)	31.4	(13.5)	32.3*	(12.5)
Econoviews International, Inc.	86–92	31.9	(11.1)	33.9*	(10.9)	30.0	(11.1)
Deutsche Banc	96–01	31.9	(18.1)	31.9	(17.7)	32.0	(18.9)
Morris Cohen & Associates	86–96	31.9*	(12.6)	38.4***	(9.9)	24.6	(11.3)
Morgan Stanley	97–01	34.0*	(14.5)	34.1*	(14.6)	33.7*	(14.5)
Ford Motor Company	96–01	34.2*	(13.1)	35.0**	(12.4)	33.1*	(14.3)
Business Economics, Inc.	86–89	40.7**	(6.7)	41.4**	(5.0)	39.9**	(8.1)

Note: Numbers in parentheses are standard deviations. *, **, and *** represent significance at the 90 percent, 95 percent, and 99 percent confidence levels, respectively.

FIGURE 4

Macroeconomic Advisors and Blue Chip Consensus Forecast Ranks

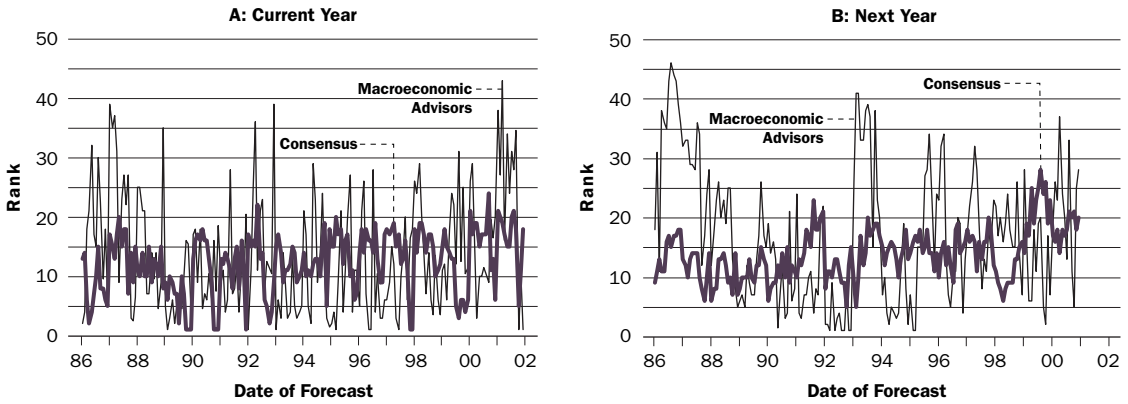


TABLE 4

Average of "Super Consensus" Scores and Ranks, 1992–2001

	Over Prior Year	Over Prior 2 Years	Over Prior 3 Years	Over Prior 4 Years	Over Prior 5 Years
			Scores		
Best forecaster	54.7 (31.1)	58.9 (28.5)	63.9 (26.1)	63.8 (26.8)	62.3 (27.0)
3 best forecasters	60.4 (27.1)	65.4 (25.2)	66.0 (24.6)	66.2 (24.8)	65.6 (23.9)
5 best forecasters	63.6 (25.3)	66.7 (23.9)	66.7 (23.5)	66.6 (23.3)	66.7 (23.3)
10 best forecasters	66.5 (23.6)	67.4 (22.9)	66.9 (22.8)	67.3 (23.5)	67.2 (22.5)
15 best forecasters	67.0 (23.3)	67.4 (23.3)	67.2 (23.0)	67.1 (23.2)	67.1 (22.7)
25 best forecasters	66.8 (23.0)	66.8 (22.8)	67.2 (22.9)	66.9 (22.9)	67.0 (22.5)
Consensus forecast	66.3 (23.0)	66.3 (23.0)	66.3 (23.0)	66.3 (23.0)	66.3 (23.0)
			Ranks		
Best forecaster	21.8 (15.5)	20.5 (13.9)	17.1 (11.6)	17.3 (11.7)	17.4 (11.5)
3 best forecasters	18.6 (12.8)	15.7 (9.7)	15.3 (9.8)	14.9 (9.2)	15.0 (8.3)
5 best forecasters	16.7 (10.8)	14.3 (8.1)	14.4 (8.9)	14.2 (7.5)	13.9 (7.3)
10 best forecasters	14.1 (7.4)	13.4 (6.0)	13.9 (6.4)	13.5 (6.0)	13.7 (5.8)
15 best forecasters	13.7 (6.4)	13.3 (5.5)	13.6 (5.3)	13.7 (5.3)	13.8 (5.2)
25 best forecasters	13.9 (5.4)	13.8 (4.7)	13.4 (4.7)	13.5 (4.8)	13.7 (4.7)
Consensus forecast	14.2 (4.7)	14.2 (4.7)	14.2 (4.7)	14.2 (4.7)	14.2 (4.7)

Note: Standard deviations are in parentheses.

APPENDIX

Data Description

Gross domestic product: 1986–95, not chained; 1996–current, chained 1996 dollars. (Note that data are revised only through March after the forecast year.) Source: U.S. Department of Commerce, Bureau of Economic Analysis, *Gross Domestic Product*, table 3.

Consumer price index: CPI-U is all urban consumers. Source: U.S. Department of Labor, Bureau of Labor Statistics, *Consumer Price Index*.

Unemployment rate: Unemployment rate (all workers). Source: U.S. Department of Labor, Bureau of Labor Statistics, *Employment Situation*.

Three-month Treasury bill: Three-month Treasury bills, secondary market (monthly average). Source: Board of Governors of the Federal Reserve System, "Selected Interest Rates," Release H.15.

Corporate bonds—1986–95: Moody's Corporate Bond Yield, Aaa (monthly average). Source: Moody's Investors Service, Inc.

Ten-year Treasury note—1996–current: Ten-year Treasury note yield at constant maturity (monthly average). Source: Board of Governors of the Federal Reserve System, "Selected Interest Rates," Release H.15.

REFERENCES

Bates, John M., and Clive W.J. Granger. 1969. The combination of forecasts. *Operational Research Quarterly* 20, no. 4:451–68.

Clemen, Robert T. 1989. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting* 5, no. 4:559–83.

De Menezes, Lilian M., Derek W. Bunn, and James Taylor. 2000. Review of guidelines for the use of combined forecasts. *European Journal of Operational Research* 120, no. 1:190–204.

Eisenbeis, Robert, Daniel Waggoner, and Tao Zha. 2002. Evaluating *Wall Street Journal* survey forecasters: A multivariate approach. *Business Economics* 37 (July): 11–21.

Newbold, Paul, and Clive W.J. Granger. 1974. Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society*, ser. A, 137, pt. 2:131–46.

Ottaviani, Marco, and Peter Norman Sorensen. 2003. The strategy of professional forecasting. Unpublished working paper.

Zarnowitz, Victor, and Louis A. Lambros. 1987. Consensus and uncertainty in economic prediction. *Journal of Political Economy* 95 (June): 591–621.