CHRISTOPHER A. SIMS
Princeton University

The Role of Models and Probabilities in the Monetary Policy Process

This is a paper on the way data relate to decisionmaking in central banks. One component of the paper is based on a series of interviews with staff members and a few policy committee members of four central banks: the Swedish Riksbank, the European Central Bank (ECB), the Bank of England, and the U.S. Federal Reserve. These interviews focused on the policy process and sought to determine how forecasts were made, how uncertainty was characterized and handled, and what role formal economic models played in the process at each central bank.

In each of the four central banks, "subjective" forecasting, based on data analysis by sectoral "experts," plays an important role. At the Federal Reserve, a seventeen-year record of model-based forecasts can be compared with a longer record of subjective forecasts, and a second component of this paper is an analysis of these records.

Two of the central banks—the Riksbank and the Bank of England—have explicit inflation-targeting policies that require them to set quantitative targets for inflation and to publish, several times a year, their forecasts of inflation. A third component of the paper discusses the effects of such a policy regime on the policy process and on the role of models within it.

The large models in use in central banks today grew out of a first generation of large models that were thought to be founded on the statistical theory of simultaneous-equations models. Today's large models have

This research was supported in part by the Princeton Center for Economic Policy Studies.

completely lost their connection to this theory, or indeed to any other probability-based theory of inference. The models are now fit to data by ad hoc procedures that have no grounding in statistical theory. A fourth component of the paper discusses how inference using these models reached this state and why academic econometrics has had so little impact in correcting it. Despite their failure to provide better forecasts, their lack of a firm statistical foundation, and the weaknesses in their underlying economic theory, the large models play an important role in the policy process. A final component of the paper discusses what this role is and how the model's performance in it might be improved.

The Policy Process

At all four central banks the policy process runs in a regular cycle; that cycle is quarterly in frequency at all except the Federal Reserve, where it is keyed to the meetings of the Federal Open Market Committee (FOMC), which take place roughly every six weeks. Each central bank has a primary macroeconomic model but uses other models as well. The primary models are the ones used to construct projections of alternative scenarios, conditional on various assumptions about future disturbances or policies or on various assumptions about the current state of the economy. Where there is feedback between models and subjective forecasts, it is generally through the primary model.

The primary models have some strong similarities. The ECB's model contains about fifteen behavioral equations, the Bank of England's twenty-one, the Riksbank's twenty-seven, and the Federal Reserve's about forty.[1] Each has at least some expectational components, with the Federal Reserve and Riksbank models the most complete in this respect. Those central banks whose models are less forward-looking describe

[1] The ECB model equations are laid out in Fagan, Henry, and Mestre (2001), and the Bank of England's in Quinn (2000). The Riksbank model is said to be nearly identical to the QPM model of the Bank of Canada, which is described in Poloz, Rose, and Tetlow (1994), Black and others (1994), and Coletti and others (1996). The Federal Reserve's model is described on a World Wide Web site that provides linked equation descriptions and a set of explanatory discussion papers. This material was made available to me for the research underlying this paper and is available from the Federal Reserve Board to other researchers on request.

them somewhat apologetically, suggesting that they are working on including more forward-looking behavior.

The Riksbank and the Bank of England have publicly described “suites” of models of various types, including vector autoregressive (VAR) models, smaller macroeconomic models, and optimizing models. Some of these models produce regular forecasts that are seen by those involved in the policy process, but at both central banks none except the primary model has a regular, well-defined role. The other central banks also have secondary models with some informal impact on the policy process.

Each policy round proceeds through a number of meetings, through which a forecast is arrived at iteratively, but the number of meetings and the way discussions are ordered vary. At the Riksbank there is a start-up meeting, at which forecasts from two large models are presented, followed by another meeting at which the sectoral experts (of which there are fourteen, with nearly everyone on the monetary policy staff responsible for at least one sector) present their views and relate them to the model. At a third meeting the staff’s report is put together. Following that meeting, an editorial committee consisting of three to five people rewrites the report into a form suitable for issue as a policy statement by the board. At this stage and earlier there is some feedback from the policy board, intended to avoid sharp divergences between its views and those of the staff.

At the Bank of England each policy round involves six or seven meetings—fewer than until recently—and some policy board members attend the meetings from the earliest stages. This may reflect the unusually high proportion of graduate-trained economists on the policy board (the Monetary Policy Committee, or MPC). All of the discussion of projections and policy choices occurs within the framework of the primary model, known as the Macroeconomic Model (MM). When a section of the model is overridden, that is done through residual adjustments, and as a result large residuals become a check on such model revisions.

At the ECB participation in the process is limited primarily to the staff until the late stages. The process begins with the gathering of projections and assessments of current conditions from the European national central banks. As at other central banks, sectoral experts play a major role, and the primary model (the Area-Wide Model, or AWM) is used to generate residuals corresponding to the expert forecasts. Twice a year a more elaborate process is undertaken, in which European national central bank staff are represented on the forecast committee and forecasts are developed through iterations with the national central banks as well as between sectoral experts.

At the Board of Governors of the Federal Reserve, the policy process (known as the Green Book process) begins with meetings among a group of about four staff members to set the "top line": forecast values for GDP growth and for certain key financial variables, including the federal funds rate. At the next stage the sectoral experts generate forecasts for the variables for which they are responsible. Their forecasts are fed through the primary macroeconomic model (called FRB/US) to generate residuals, and the results of this exercise are considered at a subsequent meeting. Feedback can occur between model forecasts and subjective forecasts and vice versa, as well as back to the top line numbers.

Federal Reserve staff emphasized to me (and it may be true at the other central banks as well) that the expert subjective forecasters have econometric input well beyond that in the primary model residuals. The sectoral experts generally have one or more small econometric models of their own sectors, and often these are more sophisticated than the corresponding equations in FRB/US. The Federal Reserve has an explicit policy of maintaining the forecast as purely a staff forecast, not allowing any policy board participation in the meetings that go into forecast preparation.

Each of the central banks prepares more than just a single forecast. The Federal Reserve probably does the most along these lines: recent Green Books show as many as a dozen potential time paths for the economy, corresponding to varying assumptions. The staff see these scenarios as a concrete way of indicating what uncertainty there may be about their forecast, despite the absence (most but not all of the time) of stochastic simulations in their analysis. The Bank of England also formulates forecasts reflecting alternative scenarios as a way of indicating uncertainty. Its inflation reports regularly publish forecasts reflecting one or more main minority views on the MPC and a forecast conditioned on the time path of interest rates implied by current conditions in financial markets. It also publishes its main forecasts as fan charts: graphs that show, as shaded regions, regions of numerically labeled higher and lower probability for the future time path of a variable.

All the central banks discussed here except the Federal Reserve condition their forecasts on an assumption of constant interest rates. This is a source of serious analytical difficulty for the Riksbank modelers, because the model they use, the Quarterly Projection Model (QPM), was built around an assumed policy reaction function. If the interest rate is truly left constant, the model explodes. If instead it is left constant for one or two years and modeled with the reaction function thereafter, it jumps at the transition date and causes strange behavior. To avoid these problems, the Riksbank simply uses the time path of long-term interest rates generated from a model run with the reaction function in place, even though the short-term rate is set on the constant-rate path. The inflation-targeting central banks are no doubt concerned that a time path for interest rates that is not flat might, if published, be given too much weight by the markets and might be seen as a commitment by the central bank. On the other hand, the ECB, which does not publish its staff forecasts, nonetheless uses the constant-interest-rate assumption, justifying it as a response to the wishes of the policy board.

Forecasting at the Federal Reserve

How well does the Federal Reserve Board staff forecast? The conclusion of this paper, which largely matches that of Christine Romer and David Romer,[2] is that the Federal Reserve forecasts quite well indeed, especially inflation. This section goes beyond Romer and Romer by

—extending their sample, which went through 1991, to 1995 or 1996

—considering data on the Federal Reserve's internal, model-based forecasts as well as data on their Green Book forecasts

—applying some analytical methods that may give additional insight into the nature of the Federal Reserve's forecasting advantage, and

—speculating on the implications of these results, in part based on my interviews with the staff, along lines that only partially match the Romers' discussion.

The Data

Before each meeting of the FOMC, the staff prepares a forecast, which is presented in the Green Book. This forecast is labeled "judgmental." In September 1995, to cite one example, it included forecasts for fifty-three

[2] Romer and Romer (2000).

variables, although the list has fluctuated in length, exceeding eighty variables in the early 1980s. The forecasts include estimates for the current quarter and projections for future quarters; over the period 1979–95 the time span of these forecasts varied from four to nine quarters. (Information about more recent forecasts is unavailable, because Green Book forecasts remain undisclosed for five years.) A "model-based" forecast is prepared at the same time. Until 1995 these forecasts were based on the MPS model, an economy-wide model originally developed as an academic collaboration but afterward maintained by the Federal Reserve Board staff.[3] Since 1995 the model used for these forecasts has been the new FRB/US model, created within the Federal Reserve. These model forecasts are archived in machine-readable form and were made available to me for this study. Their public use is, as I understand it, restricted only by the same five-year disclosure rule that governs the Green Book forecasts. The data for the MPS model forecasts that I have used, and for the FRB/US model forecasts as the five-year blackout window advances in time, will be available to researchers upon request to the Federal Reserve's research department.

This paper also considers forecasts from the Survey of Professional Forecasters (SPF). This survey, begun in 1968 as a project of the American Statistical Association and the National Bureau of Economic Research, was taken over in 1990 by the Federal Reserve Bank of Philadelphia. Data from this survey are available at the Philadelphia Fed's website.

Because some of the analyses in this section are greatly simplified by having data of uniform frequency, all the data have been converted to quarterly form. The SPF is quarterly to start with. FOMC meetings occur at least once each quarter, but with nonuniform timing within the quarter. Dean Croushore of the Federal Reserve Bank of Philadelphia has created and published on the Philadelphia Fed's website a quarterly series of Green Book forecasts, constructed by taking the FOMC meeting date closest to the middle of each quarter. Those data are used in this study. The MPS model forecasts have been put in quarterly form by matching their dates to the Croushore quarterly FOMC dates.[4]

[3] The abbreviation stands for MIT-Penn-SSRC, the three institutions that collaborated in its development (the Massachusetts Institute of Technology, the University of Pennsylvania, and the Social Sciences Research Council).

[4] The Romers chose to convert their data to monthly form instead, ending up with data sets with nonuniform timing. For their regression analyses this created no great analytical

The "actual" values used to construct forecast errors in this study are real GDP growth and inflation measured using the GDP deflator; values for the latter are those in the most recently available chain-weighted data. The Romers instead used the second revision, which appears with about a one-quarter delay. Subsequent revisions are often substantial, as are the differences between chain-weighted and fixed-weight series. However, there is an argument for targeting a near-term revision as "actual," as the Romers did. Interest in the forecasts and their influence on decisions is greatest in the months immediately surrounding the forecasts; hence errors as perceived at that time are probably closest to what enters the forecasters' own loss functions. It also seems unfair to penalize forecasters for "errors" that arise because an "actual" series is based on a different accounting concept than the series the forecasters were in fact projecting. On the other hand, the Romers have already considered actual values defined this way, and there is insight to be gained from a different approach.

The most recent revisions should, after all, be the best estimates of the actual historical path of the economy. Arguably, one should not penalize a forecaster for failing to forecast a recession that disappears in later revised data, or for predicting growth that actually occurred but was not recognized until the data were revised a year or two later. The chain-weighted data, although not available at the time most of the forecasts considered here were made, have a claim to be more accurate than the fixed-weight data available for most of the historical period I study. On these grounds, then, it is worth knowing whether analysis of forecasting performance is sensitive to whether one measures "actual" outcomes as second revisions or as the latest revisions of the most recently developed accounting concepts. That this study finds results very similar to those of the Romers supports the comforting conclusion that sustained patterns of forecast accuracy or inaccuracy are not sensitive to the details of data definitions.

Characterizing Inflation Forecast Accuracy

Table 1 shows the root mean square errors of four inflation forecasts over the period (forecasts made in 1979–95) for which all four are avail-

difficulty, and it let them preserve more of the information in the original data set. This paper's VAR-based analysis of the marginal contribution of Green Book forecasts in the presence of other variables would be complicated by nonuniform time intervals in the data.

Table 1. Root Mean Square Errors of Naïve, Survey-, and Model-Based Inflation Forecasts, 1979–95

Percentage points

	Quarters after the current quarter				
Forecast	*0*	*1*	*2*	*3*	*4*
Naïve[a]	0.00	0.94	1.15	1.14	1.35
SPF[b]	0.80	1.02	1.22	1.41	1.54
Green Book[c]	0.96	0.91	0.92	0.99	1.16
MPS[d]	1.10	1.08	1.16	1.10	1.24

Source: Author's calculations.

a. Forecast is the same rate of inflation as in the current quarter.

b. Average of forecasts reported by the Survey of Professional Forecasters.

c. Forecast prepared by the staff of the Board of Governors of the Federal Reserve for use in deliberations of the Federal Open Market Committee.

d. Forecasting model developed by researchers at the Massachusetts Institute of Technology, the University of Pennsylvania, and the Social Sciences Research Council for use by the Federal Reserve.

able. In addition to the SPF, Green Book, and MPS model forecasts, a "naïve" forecast (of no change in inflation from the current period) is presented. As the first column of the table shows, the forecasts made in real time have substantial error even in determining current-quarter inflation, for which data are available only with a delay. The naïve forecasts are therefore not naïve at all for the current quarter and are probably an unrealistic standard even one quarter ahead, because of the information advantage they reflect. At two or more quarters ahead, the Green Book and MPS forecasts, but not the SPF forecast, are better than the corresponding naïve forecast. The best nonnaïve forecast, uniformly for horizons of one quarter through four quarters, is the Green Book forecast. On the other hand, the differences between forecasts do not seem large, especially between the MPS model and the Green Book.

The similarity of the inflation forecasts is also apparent in the correlation matrices shown in table 2. Indeed, the forecasts are in general more strongly correlated among themselves than they are with the actual data. Figure 1 illustrates the same point, showing the four-quarter-ahead forecasts and the actual data tracking each other closely. A similar plot for one-quarter-ahead forecasts would be even more tightly clustered.

On the other hand, when actual inflation is regressed on the forecasts as in Romer and Romer, the results are similar to theirs (table 3): the coefficients on the Green Book forecasts are large and significant, even exceeding 1 at the one-year horizon, whereas those on the other forecasts are insignificant or even negative.

Table 2. Correlations between Naïve, Survey-, and Model-Based Inflation Forecasts and Actual Inflation, 1979–95[a]

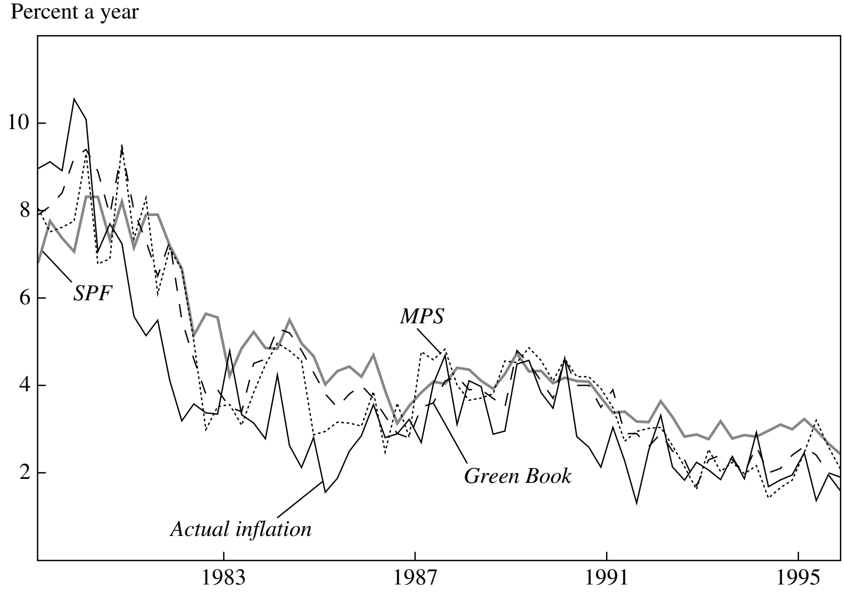
Correlation coefficients

Forecast	*SPF*	*Green Book*	*MPS*	*Actual inflation*
Four quarters ahead				
Naïve	0.9079	0.9336	0.8937	0.8324
SPF		0.9530	0.9106	0.8117
Green Book			0.9528	0.8877
MPS				0.8486
One quarter ahead				
Naïve	0.9488	0.9327	0.9091	0.9170
SPF		0.9539	0.9282	0.9212
Green Book			0.9494	0.9458
MPS				0.8963

Source: Author's calculations.

a. See table 1 for definitions. Inflation is measured by the annualized quarterly change in the logarithm of the chain-weighted GDP deflator.

Figure 1. Four-Quarter-Ahead Forecasts of Inflation, 1979–95



Source: Data sources listed in the text.

Table 3. Regressions of Actual on Forecast Inflation, 1979–95[a]

	Four-quarter-ahead forecasts		*One-quarter-ahead forecasts*	
Forecast	*Regression coefficient*	*Standard error*[b]	*Regression coefficient*	*Standard error*[b]
SPF	–0.4882	0.30683	0.2677	0.1596
Green Book	1.2564	0.32340	0.7750	0.1578
MPS	0.0444	0.23511	–0.0652	0.1231
Constant	0.3447	0.61875	–0.4221	0.2553
Summary statistics:				
R^2	0.8009		0.9750	
Standard error of the estimate	0.9751		0.7472	

Source: Author's regressions.

a. See table 1 for definitions.

b. Standard errors account for third-order moving-average, or MA(3), serial correlation.

The Romers refer to this sort of regression as measuring the "information content" of forecasts, following Ray Fair and Robert Shiller, who were probably the first to use this language to characterize this sort of regression.[5] Although the regression provides useful information if interpreted carefully, it is probably misleading to think of it as characterizing "information content." Clearly these inflation forecasts in some sense have very nearly the same "content," because they are so highly correlated.

Consider two different models of how forecasts might be related to each other and to actual outcomes. Let **f** be the vector of forecasts and y the outcome. One possible model is

$$y_t = \gamma \mathbf{f}_t + \varepsilon_t, \tag{1}$$

with the elements of the $\mathbf{f}_t$ vector independent of each other and of ε_t. Then the coefficients in the γ vector, squared, would be direct measures of the accuracy of the elements of $\mathbf{f}_t$, and they would be estimated properly by a least squares regression.

Another extreme possibility, however, is that all forecasters have noisy observations on a single "forecastable component" of y, which they may

[5] Fair and Shiller (1989).

or may not use optimally. Then if we let $\mathbf{f}^*$ denote the forecastable component of y, we have the model

$$\mathbf{f}(t) = \delta + \Lambda f^*(t) + \varepsilon(t) \tag{2}$$

$$y(t) = \varphi + \theta f^*(t) + \nu(t) \tag{3}$$

$$Var\left(\begin{bmatrix}\varepsilon(t)\\ \nu(t)\end{bmatrix}\right) = \Omega, \tag{4}$$

with Ω diagonal and f^* orthogonal to ε and ν.

In this framework the quality of a forecast is related inversely to the variance σ_i^2 of the corresponding $\varepsilon_i(t)$ and to the deviation of its λ_i coefficient (from the diagonal of the Λ matrix) from θ. It can be shown that this model implies that the estimated regression coefficients in equation 1 will all be positive and proportional to λ_i/σ_i^2. If some forecasts have very small σ_i^2 values, the relative sizes of coefficients can be extreme, then, even though the forecasts have very similar forecast error variances. Note that the coefficients are not proportional to the forecast error variances, which include a perhaps dominant contribution from the variance of ν; the coefficients are inversely proportional to the relative idiosyncratic ε_i variances, even if these are an unimportant component of overall forecast error.

Interpretation of the regression coefficients becomes even more problematic if we admit the possibility of a second component of common variation, namely, a "common error." This can be allowed for by making $\mathbf{f}^*(t)$ two-dimensional, with the second element of θ set to zero. This enables the second element of $\mathbf{f}^*(t)$ to account for similar fluctuations in the forecasts that are unrelated to the actual outcome. When there is a common component of error, the regression coefficients in models like those of table 3 can be extreme, even though the common component of error is small and very similar among the forecasts. To see this, suppose the idiosyncratic component $\varepsilon(t)$ is negligibly small, while

$$f_1(t) = f_1^*(t) + \lambda_1 f_2^*(t) \tag{5}$$

$$f_2(t) = f_1^*(t) + \lambda_2 f_2^*(t) \tag{6}$$

$$y(t) = f_1^*(t) + \nu(t). \tag{7}$$

Assuming the two components of $\mathbf{f}^*$ are uncorrelated, the coefficients of f_1 and f_2 in a regression like equation 1 or the regressions in table 3 will be $\lambda_2/(\lambda_2 - \lambda_1)$ and $-\lambda_1/(\lambda_2 - \lambda_1)$. Thus the coefficients will tend to be of opposite sign, with the difference between them *growing* as the forecasts become more similar $(\lambda_1 \rightarrow \lambda_2)$.

The coefficients of equation 1, estimated by least squares, do always retain the interpretation of being estimates of the weights in the most accurate forecast that could be constructed as a linear combination of the elements of $f(t)$. What we have seen here is that these weights have no reliable relation to individual forecast quality as measured by root mean square error. It is not a good idea, then, to limit the analysis of forecast quality to an examination of this sort of regression. It is necessary to go further in examining the correlation structure of the forecasts.

Despite its simplicity, the model of equations 2 and 3 approximates well the actual properties of the forecasts examined here. Table 4 shows estimates of this model, for one- and four-quarter-ahead forecasts. The coefficients on $\mathbf{f}^*$ in the tables have been normalized to make the coefficient in the "actual" equation equal to 1, and the constant terms have been converted to deviations from the constant term in the "actual" equation, so that they become measures of forecast bias. The model attributes the low root mean square error of the Green Book forecasts entirely to their low idiosyncratic error. Both the naïve and the MPS model forecasts have lower bias at both short and long horizons, with the lower MPS bias particularly pronounced at the one-quarter horizon.

The model fits very well at the four-quarter horizon and fairly well at the one-quarter horizon. For the four-quarter forecasts, the standard likelihood ratio (LR) χ^2 test accepts the null hypothesis that the model is correct at a marginal significance level of 0.743, if we ignore the serial correlation inherent in the overlapping forecasts. The serial correlation would almost certainly imply greater variability in the sample covariances and hence even less evidence against the null. For the one-quarter-ahead forecasts, a standard LR test of the model against an unconstrained model rejects the null at a marginal significance level of 0.013, but the Schwarz criterion, which credits simpler models for having fewer parameters and provides consistent model selection, favors the restricted model. The forecasts have slightly fat-tailed distributions (two or three residuals, out of sixty-eight, are more than 2.5 standard deviations); if we accounted explicitly for nonnormality, this would probably further weaken any evi-

Table 4. Estimates from One-Factor Model for Inflation Forecasts[a]

	Four-quarter-ahead forecast			*One-quarter-ahead forecast*		
Forecast	*Fore-castable component (f*)*	*Variance of forecast error (σ^2)*	*Constant term (bias)*	*Fore-castable component (f)*	*Variance of forecast error (σ^2)*	*Constant term (bias)*
Naïve	1.1505	0.6740	0.3738	1.0195	0.4493	0.0765
SPF	0.8235	0.2372	0.9222	0.8714	0.2028	0.4990
Green Book	1.0794	0.0094	0.6143	1.0566	0.1712	0.4906
MPS	0.9818	0.3377	0.5058	1.0609	0.5021	0.0309
Actual inflation	1.0000	0.9511	0.0000	1.0000	0.4729	0.0000
Summary statistics:						
Log likelihood		–6.3542			–6.4929	
Unconstrained log likelihood		–6.3061			–6.3868	
T for $\chi^2 = 0.05$[b]		115			53	
T for Schwarz criterion		438			99	
T of actual sample		68			68	
σ for actual inflation		2.12			2.2811	
μ for actual inflation		3.6768			3.9741	

Source: Author's regressions.

[a] Results are from estimation of the model described in equations 2, 3, and 4 in the text.

[b] T is the sample size at which the discrepancies between the theoretical and the sample covariance matrices would be statistically significant.

dence against the model. The table shows, instead of marginal significance levels, the sample size at which the conventional LR test would become "significant" at the 0.05 level and at which the Schwarz criterion would start to favor the larger model, assuming that the data moment matrices were held constant as the sample size was increased. This perhaps provides a better index of how sensitive results might be to serial correlation and nonnormality.

Since there is uncertainty both about the bias and about the variances of the forecasts, it is interesting to ask how much evidence there is against the hypothesis that the better root mean square error of the Green Book forecasts reflects only sampling error. This can be checked by fitting the bivariate mean and covariance matrix parameters for a pairing of the Green Book errors with another model's errors, with and without the constraint that, for each forecast i, $\mu_i^2 + \sigma_{ii}$ is the same, where μ_i is the bias (the mean forecast error) and σ_{ii} the forecast error variance. Table 5 shows the results of such tests. For the comparison of the Green Book with the MPS model, likelihood values with and without the equal-root-mean-square error restriction are similar, and the sample sizes T that would be required to make these differences favor rejecting the equal-root-mean-square hypothesis are correspondingly large. For the Green Book–to–SPF comparison, however, the opposite is true. The evidence for the superiority of the Green Book forecast over the SPF forecast is strong, whereas that for the superiority of the Green Book forecast over the MPS model forecast is weak.

The results here and in the Romer and Romer paper may appear to conflict sharply with those reported by Andrew Atkeson and Lee Ohanian.[6] They claim to find that econometric model forecasts are much worse than, and Green Book forecasts no better than, a simple naïve model for forecasting inflation. However, their contrasting results arise entirely from having restricted their sample to 1984–99, a period when inflation was very stable. The naïve model they consider forecasts average inflation over the next year as the average inflation rate over the preceding year. During the 1979–83 period, such a model, because it has a half-year lag during a period when inflation rose and fell rapidly, performs very badly, worse than a naïve model that uses the previous quarter's inflation. Atkeson and Ohanian also measure actual outcomes as one-year average inflation rates rather than as one-quarter inflation rates. This is not the source of their contrasting results; I have verified that their type of naïve forecast produces almost the same root mean square error as the Green Book forecasts when it is applied to forecast one-quarter inflation rates four quarters ahead over 1984–95. But it is substantially worse, at both short and long horizons, than any of the other forecasts considered in this paper when applied to the entire 1979–95 period.

Characterizing GDP Forecast Accuracy

From table 6 it is immediately clear that the advantage of the Green Book over other forecasts is much smaller for output growth than for inflation, and that the SPF forecasts look much better for this variable.

[6] Atkeson and Ohanian (2001).

Table 5. Assessing Evidence against Equal Root Mean Square Errors for Inflation

	Four-quarter-ahead forecast		*One-quarter-ahead forecast*	
Statistic	*MPS v. Green Book*	*SPF v. Green Book*	*MPS v. Green Book*	*SPF v. Green Book*
Log likelihood difference	0.0079	0.1605	0.0264	0.2837
T for $\chi^2 = 0.05$	243	12	73	7
T for Schwarz criterion	534	26	160	15
T of actual sample	68	68	68	68

Source: Author's calculations.

The forecasts are also substantially less correlated, especially at the longer horizon (table 7). The Romer-style regressions shown in table 8 still indicate a substantially larger coefficient on the Green Book than on other forecasts, but now the differences among coefficients, although large, are statistically insignificant. Table 9 shows that, despite its generally top-ranked performance, in forecasting output growth the Green Book forecast has only a statistically negligible advantage over either the SPF or the MPS model. These results might at first seem to run counter to the fact that in table 7 the correlation of the Green Book forecast with actual outcomes is higher than that of the SPF forecast with actual outcomes. But correlations ignore forecast bias, and they ignore failures of the forecast to be scaled properly. The point of the equal-root-mean-square-error test is to account for the impact of these sources of error, along with correlations. Apparently the Green Book does a respectable job at output forecasting but lacks the advantage over the MPS model and private sector forecasts that it has for inflation forecasting.

Table 6. Root Mean Square Errors of Naïve, Survey-, and Model-Based Output Growth Forecasts, 1979–95[a]

Percentage points

	Quarters after the current quarter				
Forecast	*0*	*1*	*2*	*3*	*4*
Naïve	0.00	3.61	4.09	4.39	4.58
SPF	2.46	2.93	3.09	3.37	3.12
Green Book	2.38	2.89	3.07	3.20	3.02
MPS	2.69	3.05	3.16	3.43	3.24

Source: Author's calculations.

a. See table 1 for definitions.

Table 7. Correlations between Naïve, Survey-, and Model-Based Output Growth Forecasts and Actual Output Growth, 1979–95[a]

Correlation coefficients

Forecast	*SPF*	*Green Book*	*MPS*	*Actual*
Four quarters ahead				
Naïve	–0.1444	0.2605	0.2437	0.0180
SPF		0.2871	0.2819	0.2498
Green Book			0.8230	0.4191
MPS				0.3216
One quarter ahead				
Naïve	0.5365	0.5114	0.4004	0.3850
SPF		0.8863	0.7503	0.4664
Green Book			0.8210	0.5047
MPS				0.4411

Source: Author's calculations.

a. See table 1 for definitions.

Sources of Federal Reserve Forecast Accuracy

The preceding analysis confirms the Romers' conclusion that the Green Book forecasts are very good compared with private and naïve forecasts. They have also been historically slightly better than the Federal Reserve's model-based forecasts, although the margin of superiority is statistically thin. Where does this superiority come from?

Table 8. Regressions of Actual on Forecast Output Growth, 1979–95[a]

	Four-quarter-ahead forecasts		*One-quarter-ahead forecasts*	
Forecast	*Regression coefficient*	*Standard error*[b]	*Regression coefficient*	*Standard error*[b]
SPF	0.4853	0.4130	0.3289	0.4664
Green Book	1.4133	0.6112	0.4744	0.4813
MPS	–0.1857	0.4054	0.1224	0.3057
Constant	–1.5002	1.3170	0.7366	0.5523
Summary statistics:				
R^2	0.1967		0.2626	
Standard error of the estimate	3.0278		2.8802	

Source: Author's regressions.

a. See table 1 for definitions.

b. Standard errors account for MA(3) serial correlation.

Table 9. Assessing Evidence against Equal Root Mean Square Errors for Output Growth

	Four-quarter-ahead forecast		*One-quarter-ahead forecast*	
Statistic	*MPS v. Green Book*	*SPF v. Green Book*	*MPS v. Green Book*	*SPF v. Green Book*
Log likelihood difference	0.024094	0.009275	0.010729	0.000000
T for $\chi^2 = 0.05$	80	207	179	∞
T for Schwarz criterion	175	455	393	∞
T of actual sample	68	68	68	68

Source: Author's calculations.

Some evidence on this question emerges from embedding the Green Book inflation forecasts in VAR models. I consider three somewhat extreme hypotheses:

—that the Federal Reserve is simply making better use than other forecasters of the same collection of aggregate time series available to all

—that its forecasting advantage comes entirely from its knowledge, unavailable from published time series, of its own likely policy actions, or

—that the Fed is simply collecting better detailed information about price developments, so that if other forecasters had knowledge of actual inflation one quarter ahead, its forecasts would not be useful to them.

One can formulate each of these possibilities as restrictions on a VAR model; the last two gain some support from the data.

The first hypothesis suggests that if Federal Reserve forecast data are added to a VAR containing a list of standard quarterly variables known to be useful in forecasting, the Federal Reserve forecasts should not contribute substantially to the VAR's fit. Of course, because the VAR uses ex post data for quarter *t* in constructing forecasts for quarter $t + 1$, it has an unfair advantage over the Federal Reserve forecasts, which are made at *t* without even preliminary data on many values for that quarter. If it turned out that Federal Reserve forecasts are indeed insignificant contributors to a VAR, it would be necessary to take careful account of this bias, but in fact the result seems to point in the opposite direction. The Green Book forecasts make substantial contributions to the fit of a standard quarterly VAR, as table 10 shows. The coefficients are highly significant in two of the five equations (those for the GDP deflator and the federal funds rate); the $\chi^2(5)$ statistic computed from the coefficient estimates and their esti-

Table 10. Estimates of Vector Autoregressions Including One-Quarter-Ahead Green Book Inflation Forecasts as an Explanatory Variable[a]

Dependent variable[b]	*Coefficient on Green Book forecast*	*Standard error*	*t statistic*
GNP or GDP	–0.2558	0.1585	–1.61
GNP or GDP deflator	0.1852	0.0735	2.52
Federal funds rate	0.7145	0.2279	3.13
Commodity price index	–0.5663	2.2832	–0.25
Interest rate on three-year Treasuries	0.0481	0.6836	0.07
Summary statistics:			
χ^2 (5)		23.7	
Probability (p) that Green Book forecast improves fit		0.72	

Source: Author's regressions.

a. The VAR used four lags on each of its five dependent variables and included the previous period's one-quarter-ahead Green Book forecast. The restricted and the unrestricted VARs were estimated using a combination of a "Minnesota prior" symmetric across own and other variables with "unit root" and "cointegration" dummy observations. Parameters were $\lambda = 5$, $\mu = 2$, $\xi = 3$, $\theta = 1.5$. See appendix A for details. Green book forecasts were for the GNP deflator until December 1991, and for the GDP deflator thereafter.

b. Sample period is 1975:3 to 1997:1.

mated covariance matrix suggests that correlations among the coefficients strengthen the probability that they are nonzero.[7] Table 10 shows that the posterior odds calculated from the standard priors and equal prior probabilities on the models favor inclusion of the Green Book forecasts in the VAR with 5 to 2 odds. This odds ratio is probably the best measure of the strength of the evidence. It is not decisive in itself, but because the use of ex post current data puts the Green Book forecasts at a disadvantage, the fact that the odds ratio favors the Green Book forecasts is a rather strong indication that they contain information not available in contemporaneous values of the variables in the VAR.

[7] Non-Bayesian hypothesis tests at a conventional significance level, like the likelihood ratio test in this paragraph, are not reliable guides to choice between smaller and larger models, if only because they will not settle firmly on the smaller model when it is correct, no matter how large the sample. The Schwarz criterion is widely used instead of, or as a supplement to, likelihood ratio tests, because it does converge to the smaller model when the smaller model is correct. A posterior odds ratio uses the prior probability over the parameters, together with the model's probability over the data, conditional on the parameters, to form an overall probability distribution for the data. It then compares models by looking at how likely the observed data are from the perspective of each model's probability distribution for the data. The Schwarz criterion is derived as an asymptotic approximation to the log posterior odds ratio. A posterior odds ratio, like the Schwarz criterion, takes systematic account of the number of parameters in a model, and hence avoids being misled by overfitting.

Table 11. Estimates of Vector Autoregressions Including One-Quarter-Ahead Green Book Inflation Forecasts and Current Inflation as Explanatory Variables[a]

Dependent variable	*Coefficient on Green Book forecast*	*Standard error*	*t statistic*
GNP or GDP	–0.3609	0.2148	–1.68
Federal funds rate	0.5600	0.3089	1.81
Commodity price index	–3.3499	3.0942	–1.08
Interest rate on three-year Treasuries	–0.4064	0.9264	–0.44
Summary statistics:			
χ^2 (4)		10.02	
Probability (p) that Green Book forecast improves fit		0.2024	

Source: Author's regressions.

a. See table 10 for a description of the regressions.

If the Green Book forecasts were simply a better forecast of inflation, with the forecast errors themselves having no influence on the economy, including actual future inflation on the right-hand side of the VAR should make the Green Book forecasts insignificant in all equations. As can be seen from table 11, however, the evidence on this is mixed. The individual coefficients on the Green Book forecasts in the equations (except the price equation itself) are all less than 2, and when considered jointly they are just barely significant at the 0.05 level. Posterior odds favor the model without Green Book forecasts by about 4 to 1, which is not decisive.

If the Green Book forecasts were better solely because they reflected greater knowledge of Federal Reserve policy intentions and misperceptions, one might expect that once the actual future federal funds rate is allowed on the right-hand side of the VAR regressions, the contribution of the Green Book forecast to the fit should disappear. The evidence here is quite similar to what emerges when current inflation is included in the system (table 12). The individual *t* statistics on the Green Book forecasts are all less than 1, the joint χ^2 statistic is at the margin of significance (this time just below the 0.05 level rather than just above), and the posterior odds favor the model that excludes the Green Book forecasts, this time by about 10 to 1, which is approaching the decisive range.

Note that in table 10 the coefficient on the Green Book forecast in the federal funds rate equation is strongly positive, indeed insignificantly different from 1. This means that even when several lagged values of

Table 12. Estimates of Vector Autoregressions Including One-Quarter-Ahead Green Book Inflation Forecasts and the Federal Funds Rate as Explanatory Variables[a]

Dependent variable	*Coefficient on Green Book forecast*	*Standard error*	*t statistic*
GNP or GDP	–0.3011	0.1710	–1.76
GNP or GDP deflator	0.3026	0.2258	1.34
Commodity price index	–0.4614	2.4671	–0.19
Interest rate on three-year Treasuries	–0.1947	0.7364	–0.26
Summary statistics:			
χ^2 (4)		9.26	
Probability (p) that Green Book forecast improves fit		0.0720	

Source: Author's regressions.

a. See table 10 for a description of the regressions.

actual inflation are included in the regression, the Green Book forecast of next period's inflation is a strong predictor of next period's federal funds rate. This is consistent with the view that the Federal Reserve responds to its own forecasts of inflation and that its forecasts therefore contribute to the fit through their contribution to forecasting interest rates.

Despite the caveats about data availability, the pattern of these results is consistent with the view that the forecasting superiority of the Federal Reserve arises from its having an advantage in the timing of information—even with the view that this might arise entirely from the Federal Reserve having advance knowledge of its own policy intentions. The statistical results do not prove that this view is correct, but they support it as an interesting hypothesis.

The Role of Subjective Forecasting

The persistence of the system of aggregating the views of sectoral experts to generate forecasts, despite decades of work on formal quantitative models, suggests that the expert system makes a contribution that is not easily duplicated with a formal model. What is this contribution?

One hypothesis is that the models are flawed descriptions of the economy (which is certainly true) and that the expert judgment of seasoned economists allows more subtle and accurate understandings of the economy to be brought to bear. None of the economists I interviewed at the four central banks expressed this view. Instead they claimed that the subjective forecasters mainly provide a more accurate picture of the current state of the economy than a single quantitative model can easily provide. This view was stated most clearly and strongly by those actually involved in making subjective forecasts. They argued that they pay attention to a large amount of data from disparate sources, some of it nonquantitative. They also (this view was stated most clearly at the Federal Reserve) have an understanding of how disaggregated bits of data feed into the preparation of the aggregate numbers that are put forward with some delay by the statistical agencies. This gives the experts a starting point for their forecasts that is more accurate than can be obtained from a model that is based on data generated at fixed monthly or quarterly intervals and uses a slowly changing list of strictly quantitative variables. Because most economic variables are highly persistent, this advantage in the initial period of the forecast translates into a persistent advantage. However, several of those involved in subjective forecasting, at more than one central bank, expressed the view that the advantage of subjective forecasts is almost entirely in getting the current and the next quarter right. Extrapolations beyond this horizon, they felt, could be done more reliably with the large model.

The historical record of Federal Reserve forecasts, examined in the previous section, is consistent with this view. It was shown that Green Book forecasts made at date t for quarter $t + 1$ make a strong contribution to a VAR model's fit if they are introduced on the right-hand side of a VAR regression, with forecasts made at quarter t competing for explanatory power with other variables dated t. But the Green Book forecasts' contribution to fit became much weaker when either the next quarter's inflation or the next quarter's interest rate was introduced as a right-hand-side variable. Also, the fact that the MPS model forecasts are very close in accuracy to the Green Book forecasts, together with the strong feedback between model forecasts and subjective forecasts in the policy process (as previously discussed), is consistent with the view that large-model forecasts can be as good as subjective forecasts if given equally good assessments of initial conditions.

If this view is correct, it helps explain why both large-scale modeling and subjective, or expert-based, forecasting persist in all these central banks. For the foreseeable future, explicit quantitative models are going to be associated with fixed lists of variables. But, of course, the list of variables actually in use changes over time. In some cases recognition of a policy problem or of newly apparent gaps in existing data can lead to new data collection efforts. This is accounting innovation, and it is hard to see how it can fail to include a substantial subjective or expert-based component.

Unusual events—an oil crisis, an attack on the currency, a September 11, a data collection error—can create large disturbances in data series that are not best treated as simple draws from a historical distribution of random disturbances. Because such disturbances are large and have an apparent explanation, one can likely predict that their effects will have a different pattern of persistence or influence on other variables than a typical random draw of a disturbance. Analysis of such historically unusual disturbances—including the determination of whether they really are historically unusual—will inevitably involve an element of subjective judgment. That is, because they are unique or unusual, extrapolating their effects must rely on more than historical statistical patterns of variation.

A Brief and Selective History of Statistical Modeling for Macroeconomic Policy

Jan Tinbergen's early classic macroeconometric models collected equations estimated by single-equation methods.[8] Trygve Haavelmo pointed out that the resulting models implied a joint distribution for the data and that the models should therefore be estimated and assessed as multivariate models.[9] This insight led to further developments in statistical theory, computational methods, and applied macroeconomic modeling. This simultaneous-equations or "Cowles Foundation" approach to modeling perhaps reached its peak in the collaboration to develop the MPS model, in which many leading academics worked to create a large-scale model usable for forecasting and policy analysis.

As noted above in the discussion of the policy process, each of the four central banks examined here has a primary model. These models have

[8] Tinbergen (1939).

[9] Haavelmo (1944).

evolved over time, responding to similar problems and pressures, and each central bank clearly keeps track of the others' work, so that there are important similarities in the models' current forms.

The MPS model became the main econometric model in use at the Federal Reserve. It remained so for over fifteen years, slowly evolving over time. It was retired at the end of 1995 and replaced by the model known as FRB/US, developed by Federal Reserve staff. The Bank of Canada, meanwhile, used a sequence of models, called RDX1, RDX2, and RDXF, finally scrapping them for a quite different model, the QPM. That model is essentially the same as that used at the Riksbank, and it has influenced the modeling efforts of other inflation-targeting central banks around the world.

Econometrics and macroeconomics were active research areas during the 1970s, 1980s, and 1990s, and one might therefore have hoped that there would be clear progress moving from the early simultaneous-equations models to the MPS and the RDX models and finally to the current QPM and FRB/US models. But if there has been progress, it certainly has not been clear, and my own view is that, by and large, the changes in these models over time have been more regress than progress. This is not entirely the fault of the central bank researchers who have controlled the models' evolution. The model builders have tried to take account of what they perceived as modern developments in macroeconomics and data analysis. But academic research in these areas has paid very little attention to the central problems of modeling for macroeconomic policy in real time. The three main academic research themes to which the modelers have tried to respond are rational expectations (or, more broadly, dynamic stochastic equilibrium modeling), calibration, and unit-root econometrics. The research that has emerged in these areas applies only very awkwardly to policy modeling problems. The attempts of central bank modelers to apply this research to their problems have therefore tended to make matters worse, not better. Another research theme, VAR modeling, was obtaining results that should on the face of it have been more directly applicable to real-time policy modeling. However, even here there were important mismatches between what was going on in the academic literature and the needs of the policy modelers.

The following subsections describe the most serious problems faced by the modelers, with some discussion of the absence of guidance on them from academic research.

The Econometrics of Many Variables

The four central banks studied here use models with fifteen to forty behavioral equations, with several lags of each variable typically appearing in the model. The classic simultaneous-equations toolkit was entirely based on asymptotic distribution theory, assuming that sample size is very large relative to the number of parameters being estimated and the number of variables being considered. But in these models those conditions do not hold. Two-stage least squares (2SLS), the most widely used and easily implemented estimator suggested by simultaneous-equations theory, degenerates to ordinary least squares (OLS) when the number of instruments reaches the sample size, and in practice it gets close to OLS well before that. Econometric theory gave no guidance as to how to truncate the instrument list, or even as to whether it was a good idea to do so. Limited-information maximum likelihood (LIML) and full-information maximum likelihood (FIML) estimators had the reputation of being difficult to implement and unreliable, and of course they also have only asymptotic justification. If taking account of simultaneity implied using one of these estimation methods, it seemed to require a lot of work to end up with results that were arbitrary (if based on a truncated instrument list), almost the same as OLS (2SLS with all available instruments), or quirky (FIML and LIML).

The Need for Decentralization

Good real-time forecasting and policy analysis require processing very large amounts of data. Maintaining a good forecasting model of a scale relevant to policy is more than a one-person task. The response to this situation in central banks has been, as already discussed, to allocate responsibility for "sectors" (largely identified with particular variables in a model) to experts or groups of experts, each responsible for keeping up to date with the flow of current data in their area. The MPS model was generated in a decentralized process, in which sectors of the model were assigned to individual economists or groups of economists. The Bank of Canada's RDXF model differed from RDX1 and RDX2 in pushing this decentralization perhaps to its limit: each of the sectoral experts maintained separate equations of the overall model, with little attention to the properties of the resulting system of equations.[10] The result was a system that was worse than its predecessors in terms of its long-term simulation properties. A similar history has occurred at the Federal Reserve, in the development of its Global model. This model ties together the FRB/US model with models for thirty other countries. The list of countries has grown over time, with the need to take account of financial developments in various parts of the world. But, as a result, the model can no longer be solved in its "forward-looking" mode (that is, with model-consistent expectations). This is not a matter mainly of computational time; it reflects the nonexistence of a well-defined solution to the full system.

This is the direction in which the VAR literature veered furthest from policy modeling reality. The VAR literature began by questioning whether the dissection of models into distinct, manageable equations with small numbers of variables could be justified; it urged treating equations more symmetrically, focusing on properties of the whole system of equations rather than on individual equations. The VAR literature may have been right about this in principle, but even if the point is accepted, it still provides no answer as to how to proceed to model thirty or forty variables jointly in real time with the human resources available in a central bank research department. Nor does it directly answer the question of how to integrate the subjective input of experts, who are following more detailed data than are in the model, with the model's own results. The Federal Reserve Bank of Minneapolis did for a number of years maintain a VAR model on a policy-relevant scale, and this model did include a sectoral decomposition. However, the model had a nine-variable behavioral core that took no feedback from the sectoral detail, which instead worked off of the core recursively. The model could generate forecasts and policy projections with useful detail, but its structure implied that there was no role for sectoral expertise in improving the forecasts of the main aggregates.

Simultaneous-equations econometrics and rational expectations theory were equally unhelpful on this score, of course. The theory of simultaneous equations implies no meaningful distinction between "left-hand-side" and "right-hand-side" variables. Thus the practice of naming equations after variables and assigning a sectoral expert to each runs counter to the

[10] Coletti and others (1996, pp. 9–10).

theory. Rational expectations theory emphasized that exclusion restrictions limiting the list of variables in a behavioral equation were especially dubious, because of the way expectations make every part of the model relevant to behavior in other parts.

Integration of Stochastic Modeling with Decisionmaking Uncertainty

Textbook econometrics has remained almost entirely non-Bayesian, meaning that it maintains a sharp distinction between "unknown" but nonrandom "parameters," on the one hand, and random "disturbances" on the other. Only the estimators, not the parameters themselves, are random. But in decisionmaking under uncertainty, simple axiomatics, as well as the intuition of most decisionmakers, leads to thinking of everything unknown as subject to probability calculations, including parameter values and even which of a list of models is correct. Their conventional econometrics training leads central bank staff economists to think of this kind of practical odds calculation as unscientific, or as not econometric. Here are two examples.

In my discussions with Federal Reserve staff, two economists, on separate occasions, brought up the example of how the Federal Reserve wrestled during the 1990s with the question of whether the rate of productivity growth had undergone a permanent shift. They pointed out that evidence on this accumulated slowly and that even now the conclusion remained uncertain. They suggested that if they had proceeded "scientifically" (the actual word used by one of them), they would have tested the null hypothesis of no change in the productivity growth rate and treated it as true until it was rejected at a 5 percent significance level. But as a practical matter, they pointed out, policymakers were not interested in that sort of analysis. Policymakers wanted to know what the weight of the evidence was—what were the current probabilities—not whether a test of the null was passed. Furthermore, policymakers were weighing the probability of a change in the growth rate against the costs of erring in either direction—either assuming no change when there had in fact been a change, or assuming a change when in fact there had been none. That these elementary applications of the ideas of Bayesian decision theory were seen as unscientific and as in conflict with the use of econometrics is a sad commentary on the way econometrics is now being taught.

A similar theme emerges from the Bank of Canada documents describing the origins of the QPM:

> . . . at policy institutions, the balancing of type I versus type II errors of econometric inference matters less than the balancing of type I versus type II errors in policy advice. Thus, it may be better for a policy model to assume that a particular economic structure exists, even when the evidence is not overwhelming, if the costs of incorrectly assuming otherwise would be relatively high.[11]

This passage occurs as part of a justification for considering criteria for model fit that are not "econometric."

Inflation-targeting central banks have generally (and this applies to both the Riksbank and the Bank of England) published their inflation forecasts in the form of fan charts: time-series plots that show not a single line but a fan of differently shaded regions, with darker regions meant to be considered more probable than lighter ones. Policy boards are concerned to make it clear, when they publish inflation forecasts, that they do not have a firmly held single number in mind, but instead a distribution of possible outcomes. This makes it clear that the policy board has not made a mistake, or failed to deliver on a commitment, when outcomes deviate from their most likely values by about the expected absolute amount. When I realized this, my initial thought was that this must imply an increased role for stochastic economic models, which would be used to generate the distributions needed for these charts. But at the Riksbank and the Bank of England, econometric models are not used at all in preparing these charts.

There are two reasons for this. One is that these charts represent policy choices and commitments by the policy board. Their subjective judgment is an essential element in preparing the charts. Non-Bayesian approaches to econometrics have no conceptual framework for combining stochastic simulation of an econometric model with subjective judgment. Perhaps more important, everyone understands that the main source of uncertainty about forecasts generated by a model is not the disturbance terms in the model, but rather errors of estimation in the coefficients and uncertainty about whether this model, as opposed to a number of others from which forecasts are available, is closer to the truth. Combining subjective judgment, uncertainty about parameter values, and uncertainty across models with uncertainty about equation disturbance terms would be a technically

[11] Black and others (1994, p. 65).

demanding task in any case. But Bayesian thinking provides a clear conceptual starting point, whereas the classical apparatus of confidence intervals, tests, and distributions of estimators (as opposed to parameters) provides no useful conceptual framework for these issues. The central bank staff who work on these fan charts understand very well the importance of judgmental input and that they are working with a version of subjective probability as they prepare the charts. But they have not seen a practical way to make econometric models useful as they do it.[12]

A senior staff member of the Bank of England described an incident in which he was asked by a reporter, who had just viewed a fan chart for output growth, what was the probability of two successive quarters of negative output growth over the span of the forecast. The economist had no answer, because the simple process of extracting judgmental probability distributions that generates the charts applies only to the terminal values the charts display. The evolution of the probability bands over time that the charts display is generated entirely by interpolation. This is the kind of useful extension of the existing analysis that could be produced if models and methods that can handle all the sources of uncertainty and merge it with subjective judgment were available.

Modeling Policy Choice

Some difficult conceptual issues surround the use of a statistical model fit to historical data to project the effects of a disturbance of known type, including when the disturbance is a change in policy. These issues were confronted and analyzed clearly early in the literature on simultaneous equations, perhaps best by Leonid Hurwicz.[13] He explained why it is by definition essential to have a structural model in order to analyze interventions. The early simultaneous-equations modelers understood this point in principle, although the fact that it applied strongly to macroeconomic policy interventions was not widely recognized early on in econometric practice. The most common procedure was to assume that the policy variables in the estimation were exogenous, and then to model policy changes as changes in the time path of the policy variables. More recently, thanks mainly to the structural VAR literature, it has been rec-

[12] Blix and Sellin (2000) describe how such charts are prepared and address some of the underlying conceptual issues.

[13] Hurwicz (1962).

ognized that separating policy behavior from other influences on policy variables (such as the interest rate or the money stock) is a nontrivial problem. It has become common for models to include an equation characterizing monetary policy behavior as setting the short-term interest rate in response to the state of the economy. Policy changes are modeled as temporary or permanent changes in this reaction function, including, as a special case, setting time paths for the interest rate or (equivalently) for disturbances to the policy reaction function.

This is all straightforward, and practical policy modeling has never shown much confusion about it. However, many economists interpreted the rational expectations critique of econometric policy evaluation as implying that this way of modeling policy choices had deep conceptual flaws.[14] The idea is that setting time paths of policy equation disturbances does not change the unconditional joint distribution of the time series implied by the model, and that because (unconditional) expected welfare is determined by this joint distribution, nothing important is affected by such choices of random disturbances. Many economists, even on central bank staffs, claim to hold this view even today.[15] Actual policy projections are still done, however, and in the same way as before, for the most part. Policy simulations using the QPM and FRB/US can be carried out with model-consistent expectations, and considerable effort went into making these models capable of distinguishing such projections from the usual kind. However, such projections, for policies that are not actually permanent changes, do not respond to Sargent's objections. Furthermore, it has turned out, as discussed below, that projections with model-consistent expectations are not the type most commonly used in practice.

Rigor versus Fit

Most economists would agree that a policy model should ideally be derived from a theory of the behavior of economic agents who interact to

[14] This viewpoint was perhaps best expressed by Sargent (1984).

[15] This is the case despite my having explained its fallaciousness with some frequency (for example, in Sims, 1987). The basic idea is that although the unconditional distributions do not change, conditional distributions as of the current date do change, and this matters a great deal. Furthermore, the permanent change in "rule" to which Sargent would limit attention is actually only a special case of choosing "shocks" to policy. A change in rule, when it occurs, is the realization of a random variable from the point of view of a rational public.

generate an equilibrium. Models that meet this criterion, however, are generally nonlinear and difficult to solve. Those that are soluble with reasonable time and effort tend to be fairly small and to be built around a small number of sources of stochastic disturbance. For these reasons they tend not to fit the data nearly as well as more loosely restricted models, such as VARs. This situation has been a source of major concern at central banks for some years, and they have research under way to try to overcome the problems. Academic researchers, in contrast, have paid little attention to these issues. Those engaged in building dynamic stochastic general-equilibrium models (DSGEs) have tended to stick to small models, with little apparent interest in expanding to the scale needed by central banks. This probably reflects the view held by many of those most active in working with DSGEs that real-time monetary policy formation is not very important to economic welfare. Those involved in building structural VAR models have written papers aimed at academic audiences, engaging in disputes about the effects of monetary policy and deliberately leaving unspecified the detailed interpretation of the private sector components of their models.

Despite concern about this issue at the central banks, little progress has been made. The International Finance Section of the Federal Reserve's research department has a project under way to construct a DSGE model that might serve as a partial substitute for their Global model, which, as already noted, has become unwieldy. Those in charge of the project are not at all sure, however, how much of the Global model's function the new model can take over. The Bank of England is in the midst of a two-year project to construct a new primary model. After considerable thought and investigation, researchers there have concluded that they will not be able to construct a DSGE model that fits the data well enough to be used for forecasting. They plan to construct such a model nonetheless, and to append to it ad hoc stochastic elements that can bring it into contact with the data. This perceived tension between rigor and fit is discussed explicitly in the Bank of Canada's documentation for the QPM.[16]

Both at the Bank of Canada and at the Federal Reserve, the existing primary models were seen at the beginning of the 1990s as inadequate. At the Bank of Canada this reflected the results of radical decentralization and a focus on single-equation fit, which had produced what was seen as

[16] Coletti and others (1996).

excessive fluctuation in the form of the model and made the model sometimes misbehave when used for longer-term projections. I am not sure whether these considerations were as important at the Federal Reserve. However, at both central banks there was a desire to make the model more modern by introducing forward-looking elements, so that a distinction between anticipated and unanticipated policy changes was formally possible.

The new models, the QPM and FRB/US, have a tiered structure. In the first tier, long-run, static relationships among variables are postulated. In the QPM this tier is based on an overlapping-generations growth model whose parameters are set by calibration. In practice this means simply that no measures of uncertainty are attached to parameter estimates. The parameter estimates emerge from a mixture of subjective prior information and informal matching of model properties to some summary statistics generated from the data. In FRB/US the static relationships are in many cases generated from static regression equations. These are labeled "cointegrating equations," which allows invocation of asymptotic theory that implies that uncertainty in their coefficients is negligible, and thereby also justifies estimating them by straightforward methods, independent of the second tier's complicated dynamics.

The second tier of the models describes adjustment to the long-run equilibrium. In almost every behavioral equation there is a single designated left-hand-side variable, and a "target" for it is generated from the first-tier static relations. An equation is then estimated describing the dynamics of adjustment of the left-hand-side variable to its target value.

This breaking of the model into static and dynamic tiers is common to all four of the models considered here in detail. It can be seen as a reaction to the tendency of models built up from decentralized sectors to display inconvenient long-run simulation properties. However, whether via calibration ideology or via cointegration asymptotic theory, it also insulates the long-run properties of the model from any serious interaction with the data.

The QPM and FRB/US models introduce widespread expectational dynamics through a standard mechanism. The left-hand-side variable is assumed to be determined as the outcome of an optimization problem in which the variable tracks its target value, subject to adjustment costs that depend on squared differences in the variable of some order up to k. Such an optimization problem implies that the current value of the variable will be determined by lagged and by expected future values of the target, with symmetry restrictions connecting the pattern of coefficients on future and on past values.

This approach to equation specification does introduce expectational terms pervasively in the model, but it does not respond convincingly to the critiques of the older simultaneous-equations models from the rational expectations and real business cycle perspectives. Those critiques emphasized the interdependence of model specification across equations and the derivation of dynamics and steady states from the same internally consistent, multiple-equation, equilibrium model. That research program does not lead to models that are collections of single-equation, adjustment-to-target specifications. A clear exposition of how this approach works in the investment sector of FRB/US is presented by Michael Kiley.[17] The part of the model for this sector explains, with separate equations, demand for four kinds of investment goods: high-technology equipment, other equipment, inventories, and nonresidential structures. The possibility that these four highly interrelated variables might have dynamics that interact was apparently not considered. Equation estimates are presented separately, with no check on whether the implied restrictions on the cross-variable dynamics are consistent with the data.

In the QPM the dynamic equations in the second tier are calibrated, just as are the steady-state equations. The model is therefore probably unreliable for forecasting. FRB/US, on the other hand, is fit, largely by OLS, equation by equation. OLS estimation is possible because the expected future values in the dynamic equations are replaced by forecasts from a reduced-form VAR. In principle it is possible that simultaneity would make the resulting collection of single equations perform badly as a system. However, the use of flexible lag structures and the modest amount of simultaneity in the system apparently make it a reasonable approximation to a VAR. The process of model building included comparison of the model's impulse responses with those of a reduced-form VAR, with adjustments undertaken if the deviations were too sharp.

The Bank of England's MM and the ECB's AWM contain much more limited forward-looking components. Their dynamic equations do not have the constrained symmetry of FRB/US and QPM, so that the appearance of lags in an MM or AWM equation does not force the appearance of

[17] Kiley (2001).

corresponding lead terms. In fact, both have explicit expectational terms only in the form of expected inflation, which enters mainly the wage-price block of equations and is proxied in estimation by a weighted average of inflation in the two previous quarters.

Assessment of the Primary Models

Having given up on the statistical theory based on simultaneous equations, which seemed to be providing no guidance to models of the scale actually in use, central bank modelers have ended up back at the level of the original Tinbergen model or worse. The models they use have no claim to be probability models of the joint behavior of the data series they are meant to explain, and they are not being checked against competitors by well-defined criteria of fit. It is unlikely that any systematic improvement in the models is possible without changing this situation.

The central banks using these models have given up on any serious effort to fit the data, in large measure as a trade-off for an apparently more rigorous theoretical foundation. But the improvement in rigor is largely illusory. The single-equation "target-tracking" specifications in FRB/US and the QPM do not accord with modern DSGE theory, and they seem unlikely to give anything close to an accurate rendition of the contrast between anticipated and unanticipated policy changes. Indeed, the FRB/US model is seldom used in its "model-consistent expectations" mode. Most uses of the model are for monetary policy, with a horizon of up to two years or so. In this time frame it is not reasonable to suppose that the public would quickly perceive and act on a shift in policy behavior. The VAR forecasts are therefore likely to be more reasonable approximations to actual expectations. But further, even when accurate anticipation is more plausible, the model in rational expectations mode is said generally to imply unrealistically strong and quick responses to policy actions.

I used to argue that the big policy models, although not structural in the sense they claimed to be, were still useful summaries of statistical regularities, as they were not far from being simply big VARs with quite a few exclusion restrictions. The QPM, however, cannot be rationalized in this way, because it is entirely calibrated. FRB/US may still be a good data summary, although its tiered structure makes it difficult to assess this.

Because data are available only with a five-year delay, and because there was a six-month hiatus in the preparation of model forecasts during the 1996 transition to FRB/US from the MPS, it will not be possible to assess FRB/US's performance for a few more years, absent a change in the Federal Reserve's disclosure policy.

Directions for Improvement

Bayesian statistical inference is sometimes mistakenly thought of as a collection of "techniques" for doing the same sorts of things that can be done by other "techniques." But this is a mistake. Bayesian inference is a perspective, a way of thinking about statistical techniques, not a collection of techniques in itself. The Bayesian perspective on inference, if it were widely understood by those working on policy models, would ease the connection between modeling and decisionmaking. The non-Bayesian (sometimes imprecisely called "classical") perspective, which is more appropriate (if anywhere) in the natural sciences, imposes on itself the rule that only potentially observable data, not parameters or competing models, have probabilities attached to them. It is not possible, within the non-Bayesian perspective, to make a statement like the following: "Given the data observed so far, the probability that β lies between 1.2 and 2.7 is 0.95," or "Given the data observed so far, and taking account of our uncertainty about model parameters, the probability that next quarter's GDP growth rate will be between 1.1 percent and 2.1 percent is 0.95," or "Given the data observed so far, the probabilities of each of the three models being the correct one are 0.2, 0.1, and 0.7, respectively." Economists who have been well trained in the non-Bayesian perspective know this and often claim not to be bothered by it. But when a decisionmaker, confronted with results from three models that conflict, asks what the data imply about uncertainties across the models, he or she does not want to be told that no probability weights can be given for the models. Weighting uncertain prospects to compare the expected consequences of different courses of action is the essence of decisionmaking.

Because the need for such probabilities, conditional on observed data, is so clear, non-Bayesian statistics does attempt to produce substitutes for them. Confidence intervals, which are not probability intervals conditional on data, are often interpreted as if they were. There is a classical literature on "model selection," but it does not yield probabilities on models and has produced a bewildering variety of procedures. There are ways, from a non-Bayesian perspective, to create hybrids of confidence intervals and probability intervals that in some sense incorporate parameter uncertainty into measures of forecast uncertainty, but this is not the same thing as probability intervals conditioned on the data.

Since these substitutes are often in practice interpreted as if they were probability statements conditioned on the data, the distinction between the Bayesian and the non-Bayesian perspective is sometimes seen as philosophical hairsplitting. But there are some real costs to the persistence of non-Bayesian thinking among econometricians. There is an important difference, for example, between the situation where the data do not discriminate sharply among several competing models but one fits slightly better, and the situation where a single model stands out as best fitting. Econometric analysis should be able to give quantitative guidance as to which type of situation one is in, rather than insisting on the impropriety of putting probabilities on models. Inference about unit roots and cointegration is, from a non-Bayesian perspective, extremely complex—so complex that the academic literature provides little practical guidance about how to handle such phenomena in models of the scale of central bank primary models. The literature has also emphasized asymptotic results that in many cases appear to justify multistage inference, "testing" for unit roots and estimating cointegrating relationships, and then treating the results of this first stage of inference as if they were free of uncertainty. But in actual, finite samples the uncertainty surrounding these first-stage inferences is often high. Ignoring it produces unbelievable results. Despite not applying unit-root theory at the full system level, because of its impracticality, several of the central bank models do apply it in exactly this multistage way, equation by equation. A Bayesian perspective would allow setting aside the multistage complexity of this inference and thereby would allow a more accurate and believable characterization of uncertainty about low-frequency aspects of the model. Analysis of the possibility that a model's parameters have shifted at discrete points in time is another situation, central to the use of models in decisionmaking, that is much more straightforward to analyze from a Bayesian than from a non-Bayesian perspective.

The Bayesian perspective recognizes that all decisionmaking depends on a decisionmaker's judgment as well as on inference from the data, and it formalizes the interaction of judgment and data, in contrast to the non-Bayesian perspective, which avoids the subject. When the role of judgment is kept under the table, it becomes more difficult to discuss it and more difficult to recognize bad judgment.

But the most persuasive argument for the Bayesian perspective is its increasing ability to make apparently intractable inference problems tractable. Within the last five years or so, economists have become aware of Markov-Chain Monte Carlo methods for Bayesian analysis of econometric models. These methods are being applied rather widely, particularly in finance, and as economists begin to see them produce insights into otherwise intractable problems, they are likely to learn how to use these methods and understand their rationale. In fact, a recent paper by Frank Smets and Raf Wouter apparently represents the first example of a DSGE that has been fit to data and produces a fit that is competitive with that of a Bayesian reduced-form VAR.[18] The paper accomplishes this in a Bayesian framework, using Markov-Chain Monte Carlo methods, and thereby produces a model that should be directly usable for realistic stochastic simulation and should be directly comparable in fit to models with different specifications. Although it explains just nine variables, it was put together by two researchers in a relatively short time. With the resources of a central bank research staff and computational equipment, the same methods should work on models of the scale of today's central bank primary models. On the face of it, this makes obsolete the widespread belief that rigorous dynamic theoretical modeling and good statistical fit are incompatible.

So the problem of conflict between rigor and fit in modeling may be on its way to resolution. Since it appears that this resolution will involve an increased awareness of Bayesian ideas, there may be progress along the way in formalizing the connection between subjective judgment and model forecasts. But the problem of decentralizing modeling effort seems likely to remain difficult for some time.

A model like that of Smets and Wouter contains both firm and household sectors, but each of these sectors generates several tightly related behavioral equations. It does not seem much more suited than a VAR to equation-by-equation decentralization. Given the nature of central bank subjective forecasting, however, it might be possible to combine variable-

[18] Smets and Wouter (2002).

by-variable decentralization of expert input with a more integrated approach to modeling. It might be worth exploring a structure in which judgmental forecasters focus almost entirely on the current and the next quarter. Their forecasts could be treated as noisy observations on the data that enter a Bayesian structural model. Model forecasts could then incorporate these observations, as well as generate measures of how implausible they are.

Conclusion

As I see it, the most important component of inflation targeting is the regular reporting of the forecasts and policy analyses by the central bank that the targeting regime entails.[19] This supports central bank policy by making it easier to preserve credibility in the face of shocks that create temporary increases in the inflation rate. By announcing a policy path and a corresponding inflation path, the central bank may be able to convince people that the inflation will end without having to generate a recession. This regular reporting of forecasts also encourages probabilistic thinking and creates a demand, as yet unsatisfied, for policy models that can generate realistic measures of uncertainty about their results.

Some apparently unnecessary barriers to transparency in monetary policy persist. Other central banks around the world are regularly reporting inflation and output forecasts without ill consequences, indeed with apparently good consequences. The Federal Reserve's Green Book forecasts of inflation are of very high quality. They could be useful to the private sector and, if published in a timely way, could contribute to the effectiveness of Federal Reserve policy actions. It seems that, at least for these forecasts, the five-year embargo should be dropped.

In the inflation-targeting countries, the internal consistency of forecasts would be improved and the level of discussion of policy elevated by switching to a practice of publishing forecasts in which inflation, output, and interest rates all appear and have been derived from a model so as to be mutually consistent. The usual objection is that this asks too much of policy boards that already have difficulty agreeing on just the current

[19] For an analysis of the value of this type of transparency that supports my views here, see Geraats (2001).

level of the policy rate. However, the task of choosing among and adjusting several proposed time paths for the policy rate does not seem much more difficult than the problem of generating a skewed fan chart distribution, which policy boards are already solving. Here again is an area where considerable progress could be had cheaply.

Finally, there are the central challenges of policy modeling. Is it realistic to hope that fully identified models can be found that allow storytelling not only about the effects of policy but also about where major shocks have originated, and that fit the data? Is it realistic to expect that economists can learn to understand the Bayesian perspective on inference and how it dovetails with decision theory? Is it practical to continue to do policy modeling in teams, while also improving the models along these lines? I have argued that the answer to the first two questions is yes, with important developments in these areas likely soon. Whether modeling can be decentralized does not have as clear an answer, although there is some reason for hope.

The academic branch of the economics profession has not been contributing much to the progress of policy modeling. Increased attention by scholars to these interesting problems might resolve fairly quickly many of the problems with central bank practice I have cited. However, academic interest in these issues has been low for years, and I am not sure how this can be changed.

APPENDIX A

VAR Priors

THE ESTIMATION OF the VAR models in this paper uses priors meant to aid in interpreting the likelihood shape by concentrating attention on the most reasonable parts of the parameter space. This is necessary in the first place because time-series models, especially models where the number of parameters is a relatively large fraction of the sample size, easily produce spuriously precise results, attributing much of the observed sample behavior to unusual initial conditions, if no prior is used. This problem is the Bayesian counterpart of the cumbersome unit-root theory in classical inference.[20]

[20] Sims (1989, 1996).

In addition, I am comparing restricted and unrestricted versions of the VAR models here, and there are well-known difficulties in deciding how properly to penalize overparameterized models that have good in-sample fits. The widely applied Schwarz criterion is one approach, and it has a Bayesian justification as a decision procedure in large samples, but its justification when unit roots are possibly present is problematic. Since there is a widely used, standardized family of priors for VARs that can easily be made proper (that is, made to integrate to 1), using this prior directly rather than the Schwarz asymptotic approximation seems like a good idea, and I have done so here.

The prior probability density function (PDF) is of the form

$$|\Sigma|^{-n-1/2}|\Sigma|^{-k/2}|X_d' X_d|^{n/2}\, e^{-1/2([y_d-(I\otimes X_d)\beta]'(\Sigma^{-1}\otimes I)[yd-(I\otimes d)\beta]}, \tag{A1}$$

where y_d and X_d are dummy right-hand-side and left-hand-side variables, k is the number of columns in X_d, n is the number of equations, and β is the vector of parameters from all equations, stacked on each other. I have omitted some factors that vary with n, but not with the data or with k.

This is just convenient notation for a prior that, conditional on Σ, is Gaussian for β and with the covariance of the β's across equations fitting a $\Sigma \otimes \Omega$ form, where $\Omega = (X_d' * X_d)^{-1}$. The marginal prior on Σ is proportional to $|\Sigma|^{-n-1/2}$. This is not a proper prior. It is the limit of an inverse-Wishart prior with n degrees of freedom as the scale matrix shrinks toward zero. A proper prior with a very small scale matrix would give nearly identical results.

This prior is in conjugate form, so that, combined with the likelihood, it can be integrated analytically to provide posterior odds on models.

The dummy observations actually used were in three sets:

—*A "cointegration" dummy*. This is a single observation in which all exogenous and all current and lagged endogenous variables in the system are set equal to their means over the period of the initial conditions. If given very high weight, this dummy forces the appearance of at least one unit root affecting all variables, in which case it also enforces zero coefficients on exogenous variables, or else it forces the system toward stationarity with a steady state equal to the initial condition means. This observation is weighted by the hyperparameter λ.

—*A set of n "unit root" dummies*. These, if weighted very strongly, force the appearance of n unit roots in the system, one in each variable.

The *i*th of these is an observation in which the *i*th variable and its lags are all set equal to initial condition means, and all other variables, including the constant term, are set to zero. These observations are weighted by the hyperparameter μ.

—*A version of the Minnesota prior*. These are $nq + r$ dummy observations, where q is the number of lags in the system. For the *j*th variable's *k*th lag, the "*x*" dummy observation is nonzero only for that lag of that variable and is set equal to $\sigma_j k^\theta$, where σ_j is the standard deviation of the changes in the variable over the period of the initial conditions. There are also dummy observations corresponding to the other right-hand-side variables, sized simply at the standard deviation of the "*x*-variable" itself over the period of the initial conditions. One can think of these as applying even to the constant term, but since the constant term has zero variance, the weight on the corresponding dummy observation is zero. This whole set of observations is given weight ζ. The y_d variable is nonzero only for the lag q of 1, in which case both the left-hand-side and the right-hand-side *y*'s are set to σ_j. This makes the prior mean emerging from these dummy observations 1 for the first own lag, and zero for all other own lags and for all other variables on the right-hand side.

The VAR results reported in the tables use $\lambda = 5$, $\mu = 2$, and $\xi = 3$. The last value is close to maximizing the posterior PDF, conditional on $(\lambda,\mu) = (5,2)$. The values of λ and μ were set based on experience with similar models. They cannot be chosen to maximize the posterior PDF without reintroducing the problem of estimates implying spuriously unusual initial conditions.[21]

[21] The family of priors from which this one is drawn is discussed in more detail in Sims (1993) and Sims and Zha (1998).

Comments and Discussion

Steven N. Durlauf: This ambitious paper tackles an extraordinarily difficult question: what is the role of formal statistical models in evaluating economic policies? In particular, the paper studies the use of such models by central banks in various capacities. Although the paper addresses a wide range of issues and provides a fair amount of qualitative description of how central banks use large-scale models, its main contributions are twofold. First, it provides an evaluation of the forecasting performance of the Federal Reserve. Second, it addresses several broad questions concerning the appropriate ways of using statistical models in the policy process. I will deal with each of these components in turn.

Sims' evaluation of the forecasting accuracy of the Federal Reserve provides some useful additions to a long-standing literature, in particular a recent paper by Christina Romer and David Romer.[1] Sims compares both the judgmental and the model-based forecasts of the Federal Reserve with two alternatives: naïve forecasts and a consensus forecast from the private sector. What is new in Sims' comparison relative to that of Romer and Romer is the attention to the relative virtues of the judgmental and the model-based forecasts. The main claims Sims makes are, first, that the Federal Reserve forecasts well, especially when forecasting inflation; second, that the informational contents of different forecasts are highly correlated, so that strong claims of superiority of one forecast over another should be treated as suspect; and third, that there does not appear to be strong evidence that the judgmental forecasts of the Federal Reserve are

[1] Romer and Romer (2000).

superior (as measured by the root mean square forecast error) to its model-based forecasts.

Although these points are well taken, the analysis succeeds less well in giving a clear understanding of the differences between the model-based forecasts and the forecasts that embody subjective judgments. One limitation is that the procedures used for forecast comparison are not well chosen if one's objective is to go beyond crude summary measures of relative forecast accuracy to an understanding of why forecasts differ. Root mean square error is certainly a sensible single summary statistic for comparing forecasts, but like all such summaries it is limited. In my view, an additional useful way of comparing two forecasts is to attempt to identify periods when the two forecasts diverge relatively sharply and compare their behavior at those times. I suspect that, during shifts across business cycle regimes, differences are larger between the Federal Reserve forecasts and the private sector forecasts than in other periods. Put differently, the fact that two forecasts are approximately equally accurate in periods when they are close to each other is not informative about their relative performance when they are far apart. Presumably what one is interested in is whether, when the differences are relatively large, one forecast performs better than the other.

Further, it would seem that a deeper evaluation of the forecast differences should address in greater detail the relationship between forecasts and their use, especially if one is interested in how subjective judgment affects forecasts. In a 1997 paper (which, curiously, Sims does not reference), David Reifschneider, David Stockton, and David Wilcox give three justifications for the use of judgmental over model-based forecasts: first, the ability of the former to use "potentially valuable information contained in monthly and weekly data" not incorporated into the model; second, the integration of "extramodel information and anecdotal evidence into the forecast"; and third, the ability to address model uncertainty: "the judgmental approach . . . enables the staff to examine a range of econometric specifications—both structural and reduced form—in producing the forecast rather than relying on a single specification enshrined in the 'staff model.'"[2] These are all plausible reasons for using judgment, and all would seem relevant to evaluating the effectiveness and value of subjective judgment in Federal Reserve forecasting. Although Sims' paper gives

[2] All quotations are taken from Reifschneider, Stockton, and Wilcox (1997, p. 17).

some attention to the question of information asymmetries, virtually none is given to these other explanations for why judgmental forecasts deviate from model-based ones. Now, some of these reasons may not be identifiable from available data, but if so, that needs to be made clear. Notice as well that Reifschneider, Stockton, and Wilcox's comment on model uncertainty suggests that some of the issues raised later in this paper are part of the Federal Reserve's mindset; more on this below.

Besides evaluating the accuracy of different forecasts, Sims also makes a number of broad arguments concerning the nature of large-scale models as currently employed by central bankers, notably the Federal Reserve. Anyone even casually acquainted with Sims' past work will not be surprised to discover that this paper offers serious criticisms of the statistical models that central banks have constructed as well as the way they are used. We are told, for example, that "The large models in use in central banks today have lost any connection to the simultaneous equations–based statistical theory once thought of as the intellectual foundation of monetary policymaking. The models are now fit to data by ad hoc procedures that have no grounding in statistical theory."

Many of these criticisms are trenchant. And I am in absolute agreement with Sims on one of his main arguments: that evaluation of policies should be explicitly decision-theoretic. Too much of empirical policy analysis in academia makes claims of the form that because something is (or is not) statistically significant at the 5 percent level, a certain policy is (or is not) justified. This sort of reasoning is flawed because the evaluation of policies requires the comparison of expected benefits and losses, and therefore it cannot in general be done without attention to the complete conditional distribution of outcomes given a certain policy as well as an explicit statement of the loss function by which these outcomes are assessed.[3]

Unfortunately, much of the paper's discussion of general issues of modeling suffers from overstated and poorly justified claims. One problem is that the paper takes a dogmatic position on the virtues of Bayesian methods, which is asserted over and over without argument. For example, the paper gives no attention to the possibility of constructing decision-theoretic approaches to data analysis using non-Bayesian ideas. I am

[3] Recent efforts to incorporate decision-theoretic reasoning into data analysis include Brock and Durlauf (2001), Dehejia (forthcoming), and Manski (2000).

thinking here of the approach to statistical decision functions developed by Abraham Wald,[4] an approach that has helped motivate recent important work by Charles Manski.[5] To be fair, these approaches have not been applied to macroeconomic policy contexts, and so it is unclear what they can accomplish.[6] The point, however, is that one can take a rigorous decision-theoretic approach without adopting the Bayesian paradigm.

Similarly, Sims' enthusiasm for Bayesian approaches also leads to a number of questionable assertions about the modeling of persistence in time series. We are told that classical unit-root methods for analyzing time series are "cumbersome," that they provide "little practical guidance" for large models, and that they rely on asymptotic results that are inapplicable (producing "unbelievable results") in finite samples. None of these claims is persuasive. If cumbersome means computationally difficult, I see no reason why this applies with less force to the Markov-Chain Monte Carlo methods necessary for Bayesian inference. The second two claims misrepresent the state of classical unit-root econometrics. That literature contains a vast number of analyses that attempt to deal with multivariate systems with unknown numbers of unit roots. A seminal paper by Peter Phillips provides a comprehensive methodology that allows one to simultaneously determine the cointegrating rank and the lag length of a system of vector autoregressions; these methods move far beyond the older ones Sims criticizes.[7] This literature has also paid enormous attention to finite-sample issues related to estimators. As a whole, then, Sims' criticisms carry very little weight.

Beyond Bayesian dogmatism, a second and more serious problem for this paper is its Bayesian utopianism. Simply put, the paper fails to address the issue of how one engages in Bayesian decision-theoretic analyses in practice. Sims writes that

> Combining subjective judgment, uncertainty about parameter values, and uncertainty across models with uncertainty about equation disturbance terms would be a technically demanding task in any case. But Bayesian thinking provides a clear conceptual starting point, whereas the classical apparatus of confi-

[4] See, for example, Wald (1950).

[5] See, for example, Manski (2000).

[6] In fairness to the Bayesian approach, Wald's version of decision theory, although it avoids the need for priors, does require introduction of something like "a game against nature" instead.

[7] Phillips (1996).

> dence intervals, tests, and distributions of estimators (as opposed to parameters) provides no useful conceptual framework for these issues.

Perhaps these problems are conceptually straightforward in a Bayesian context, but they are most certainly not operationally straightforward.

This is clearest in the context of prior information. To take one example, consider the question of subjective information. Using Bayesian language, recall that the posterior density of an object θ (such as a parameter or a future outcome) given information at F_t is always written as

$$\mu(\theta \mid F_t) = \frac{\mu(F_t \mid \theta)\mu(\theta)}{\mu(F_t)} \propto \mu(F_t \mid \theta)\mu(\theta). \tag{1}$$

Where does this expression allow for subjective judgment? The answer is via $\mu(\theta)$, which is the prior probability density associated with θ. Any available information may be incorporated here, so that, at this level of generality, the incorporation of subjective beliefs is, as Sims says, conceptually straightforward. But his claim is also question begging. The incorporation of subjective judgments in the Bayesian framework presupposes that this subjective knowledge can be expressed in terms of the probability density $\mu(\theta)$. This is a long-standing concern relative to Bayesian methods, but not an idle one. David Freedman, one of the world's leading mathematical statisticians, and such a ferocious critic of overclaiming in empirical social science as to make Sims look like an unrepentant data miner, has written

> For thirty years, I have found Bayesian statistics to be a rich source of mathematical questions. However, I no longer see it as the preferred way to do applied statistics, because I find that uncertainty can rarely be quantified as probability. The Reverend Thomas Bayes had his doubts too, which is why he allowed his essay to be published only after his own death; and the matter has been debated ever since.[8]

By no means am I arguing that the problem of translating subjective judgments into probability statements invalidates the Bayesian approach. Nor am I criticizing the use of priors in general; many of the standard criticisms are uninteresting because they are made without context. For example, if an uninformative prior "converts" the ordinary least squares estimate of a parameter into the mean of the posterior density describing

[8] Freedman (1991, pp. 356–57).

the parameter, this is clearly unobjectionable. On the specifics of prior information, one can equally well find eminent statisticians (Dennis Lindley comes to mind) who will claim that the problem of prior construction is in no way an insuperable barrier to practical Bayesians. Within the Bayesian statistics literature the issue of converting subjective beliefs into probability statements is a long-standing area of research; for example, this is what the literature on Bayesian elicitation is all about. My argument, rather, is that there exist serious difficulties in operationalizing subjective information. Without an explanation of how one is actually going to do this, it is unclear that one will do better than the admittedly ad hoc judgmental approach in the Federal Reserve forecasts. Moving from the general to the specific, I would urge the reader to ask how subjective judgments about unusual events can be made operational. Does one really believe that forecasts concerning the effects of September 11 would be appreciably different had the Federal Reserve followed Sims and used formal Bayesian language to describe the way it modified model forecasts using subjective judgment?

Related problems of prior formulation exist with respect to the modeling of model uncertainty.[9] I agree with Sims' argument that it is important to account for model uncertainty in evaluating policies. A Bayesian perspective on model uncertainty can be thought of as follows. Suppose that there is a set of possible models, M, each element of which implies a joint probability relationship between the object of interest θ and F_t. When one fails to account for model uncertainty, in the sense that a particular model $m \in M$ is assumed to be the "true" model, one in essence calculates

$$\mu(\theta \mid F_t, m). \tag{2}$$

Incorporation of model uncertainty requires that one eliminate the conditioning on the particular model m. This is again conceptually straightforward. One simply treats m as a random variable and integrates over it;[10] that is,

$$\mu(\theta \mid F_t) = \int \mu(\theta \mid F_t, m)\mu(m \mid F_t), \tag{3}$$

[9] Draper (1995) provides a very lucid discussion.

[10] Usually, the model space is assumed to be countable, in which case the integral is replaced by a sum.

where

$$(4) \qquad \mu(m \mid F_t) = \frac{\mu(F_t \mid m)\mu(m)}{\mu(F_t)} \propto \mu(F_t \mid m)\mu(m).$$

Once again we run into the issue of how to specify prior information, in this case $\mu(m)$, which embodies the prior probabilities one assigns to the elements of the model space. Little work has been done on this question. To take one example, the main place where substantial work on model uncertainty is currently being done in statistics concerns uncertainty about which variables to include in a model. For linear regressions, Adrian Raftery, David Madigan, and Jennifer Hoeting discuss methods of accounting for uncertainty in a given regression coefficient as induced by uncertainty as to what other variables to include.[11] The solution of model uncertainty of this type is to calculate equation 4 for every element of a space of models *M,* where individual elements are alternative choices of regressors. If one cares about the coefficient on a given regressor where there are *K* potential additional regressors, this means there are 2^K different models in *M*. How does one set priors for this case? The standard approach in the statistics literature is to assume that each model is treated as having a probability $p^{l(m)}$ $(1 - p)^{K-l(m)}$, where $l(m)$ is the number of regressors included in model *m.*[12] This prior in essence assumes that each variable is present in a regression with a prior probability *p* and that the presence of a given variable is independent of the presence or absence of the others. As Willard Brock and I have argued,[13] in many contexts this is clearly unsatisfactory from the perspective of economic theory. Variable selection possesses many features similar to the "red bus/blue bus" problem in discrete choice theory; that is, different regressors are economically linked with one another. Leaving aside economic objections, nothing is known about the robustness of posterior densities based upon this prior. So, although no one can object to accounting for model uncertainty in principle, the hard task is doing so in a useful fashion.

The issue of prior information is also complicated in terms of the modeling of persistence. The issue of priors and unit roots has in fact been

[11] Raftery, Madigan, and Hoeting (1997).

[12] Model averaging to account for uncertainty in variable inclusion has been applied in the context of economic growth by Brock and Durlauf (2001), Fernandez, Ley, and Steel (2001b), and Doppelhofer, Miller, and Sala-i-Martin (2000).

[13] Brock and Durlauf (2001).

highly controversial. Phillips showed that Bayesian analyses of persistence under allegedly noninformative priors in fact substantially understated the evidence in favor of unit roots in various macroeconomic data series.[14] A key message of Phillips's paper is that the interpretation of priors is very difficult in the context of time series with possible unit roots. Although in this paper Sims advocates different ways of constructing priors to deal with unit roots from those criticized by Phillips, the solution seems ad hoc and certainly has not been vetted in the econometrics literature. Perhaps his approach makes sense, but one needs to be candid in admitting that it hardly constitutes received wisdom.

Finally, the paper suffers from the problem that, despite its advocacy of decision-theoretic methods, the paper itself is ultimately not very decision-theoretic. The paper's many criticisms are made outside the context of actual central bank decisionmaking. However, central bank models have not been developed with the standards of academic journals in mind; they have been produced to facilitate policymaking in extremely complicated contexts and should be judged on that basis. Unfortunately, many of Sims' criticisms are closer to debating points than substantive indictments.

For example, Sims criticizes the way expectations are handled in the FRB/US model. The criticism is that the formation of expectations does not respect the systemwide constraints in the model. But what is the import of this and related criticisms? Is there any reason to believe that the use of these single-equation estimates of expectations is a first-order issue in terms of forecasting or in terms of policy evaluation? Sims asserts this is so, but nothing in the paper supports the claim. I would submit that without some evidence that the various defects of the large-scale models raised by Sims have some effect on policy evaluation, at least in terms of the sorts of results the model produces in policy simulations, the criticism loses much of its force.

Similarly, many of Sims' criticisms of the way central banks use models seem closer to caricatures than descriptions of actual practice. The paper places particular emphasis on model users' lack of attention to model uncertainty and other defects. This contrasts with the description of model use in Reifschneider, Stockton, and Wilcox, which shows a nuanced understanding of the limits of their own FRB/US model:

[14] Phillips (1991).

> . . . policy analysis and the evaluation of monetary policy rules rely heavily on the use of large-scale macroeconometric models. However, some constructive tension always exists between the current specifications of the models and the staff's evolving understanding of macroeconomic behavior. Consequently, even in these activities, judgment remains an important element in our analysis.[15]

Sims' criticisms are also inconsistent with his own observation that the Federal Reserve reports forecasts under alternative scenarios. The bottom line is that I believe that users have a better understanding of the limitations of their models than is suggested by the rather insulting anecdotes reported by Sims.

The difficulties in translating criticisms into practice are illustrated in the treatment of how to incorporate economic theory into models. Although Sims frequently suggests that central bank models fail to properly come to grips with economic theory, his discussion is quite weak in terms of specifics, except with regard to how expectations are handled. Yet in defending his VAR approach to forecasting, Sims states that "it is not reasonable to suppose that the public would quickly perceive and act on a shift in policy behavior. The VAR forecasts are therefore likely to be more reasonable approximations to actual expectations." It is unclear to me why the willingness to make assumptions like this places the VAR approach any closer to consistency with economic theory than the loose theorizing used to motivate the expectations specifications in the large-scale models. In fact, the question of how to best incorporate economic theory into a forecasting model is very difficult. Does economic theory suggest that such a model should embody Ricardian equivalence, as forward-looking rational behavior implies, or should it not, as the presence of liquidity constraints suggests? The answer is far from obvious and in any event requires a specification of the objectives for which the model is being used.

In conclusion, this paper falls prey to the same criticism of lack of policy relevance that it makes of the models and their uses. Issues of the use of subjective information, accounting for the model, and the like are all at some level apparently part of the Federal Reserve's actual procedures in making and assessing forecasts. The paper's criticisms thus too often fall into the area of formalism rather than substance. Policymaking at the

[15] Reifschneider, Stockton, and Wilcox (1997, pp. 21–22).

Federal Reserve can be analogized to an engineering problem. It is all well and good to point out where an engineering project takes short cuts, deviates from high theory, and the like. (Theoretical physics is similarly replete with cases of approximations, normalizations, and other shortcuts that make mathematicians cringe.) I am certainly willing to believe that, in principle, formal criticism of this type can be translated into practical suggestions that allow the models to better assist policymakers. Yet not all logical defects are equally important in practice. Without some demonstration of how various identified flaws affect the actual use of models in informing policy and providing clear operational guidance on how to improve the existing large-scale models, the many criticisms made here are unlikely to have much effect.

Jeffrey C. Fuhrer: I will begin my discussion with some general remarks about the state of monetary policy models and then turn to some specific comments on Christopher Sims' paper.

The use of models is almost surely unavoidable in policymaking, as it is in virtually every aspect of life. When I describe my job to my children, my spouse, and my noneconomist friends, they often ask me something like, "So what's with all that modeling stuff?" I usually reply, much to their delight, that *everyone* uses models, even in everyday life. For example, we all use a "model" of a car when we drive it: we conveniently abstract from the details of the internal combustion engine, the suspension and cooling systems, and the physics that links them all together with the driver, because these aspects of the vehicle are generally unimportant for everyday driving decisions. This model is so much a part of our lives that we take for granted that our means of interacting with our cars is a simplified, processed representation of a far more complex external reality. But it is through this model that we are able to use and make real-time sense of an otherwise remarkably complicated machine.

After rousing my involuntary audience from their stupor, I usually continue by pointing out that this example shares a number of features common to all models: It simplifies a more complicated reality. It therefore abstracts from features of reality that may be important in other circumstances. The way in which it simplifies reality depends critically upon the use to which the model is to be put (compare the mechanic's model of the car with the driver's). It imposes some structure on reality that both captures essential elements of that reality (where what is "essential" is gauged by the user) and makes it more "usable." And it is not limited to statistical or even to mathematical representations.

The same principles apply to economic policymakers and macroeconomic models. In fact, I do not believe there is any coherent way to conduct policy without employing some kind of model. The model that monetary policymakers actually use may not look like the models on Sims' website. It might reside entirely between the ears of the current Federal Reserve chairman. But every policymaker has some way of imposing a simplifying structure on the complex interactions between economic actors and economic policy. This structure is both the means by which the policymaker is able to interpret (that is, identify) the current economic environment and a tool for considering policy alternatives. Thus models for policymaking, of one sort or another, are almost surely unavoidable.

The question then becomes: What kind of models should we use, what kind of models *do* we use, and what and how large are the differences between them? I will focus on the formal models developed by economists for policy analysis.

In his own comments on other papers, Sims notes that a number of authors have "examined the last half century or so of US monetary policy, evaluating how good it has been, whether it has been improving, and what forces have contributed to its evolution," but he goes on to caution that ". . . the apparent lack of consensus on the sources of monetary nonneutrality, the dependence of conclusions on assumptions about these sources, and the limited attention to statistical fit in this literature make its conclusions on these issues at best tentative."[1] I think this sums the matter up nicely. In other words, the conclusions reached in this literature might well be interesting, and even useful, if only our profession had reached the stage where our models were really up to such a task.

But they are not. The gap between the ideal and the practical is large. So what *should* a model do? First, it should be based on theory that is designed to identify underlying household, firm, and government behavior—objectives and constraints, for lack of better terms—so that we can both interpret the data in terms of economic behavior and conduct welfare analysis. Second, the model should be able to pass empirical tests that determine whether the desired identification has been achieved, whether

[1] Sims (2000, p. 5).

the model fits the data (where I construe "fit" broadly), and whether the model appears to be stable in the presence of modest changes in policy regime.

What *do* the models do? Unfortunately, the models that are strongly linked to theory cannot pass the empirical tests, and the models that pass some of the empirical tests deviate markedly, and usually in ad hoc fashion, from theory. Instances of the first sort are to be found in many academic papers featuring optimizing policy models. Instances of the second sort are to be found at many central banks.

Turning first to the theory-linked models, most of the extant academic literature can be reduced to output and price-setting equations of the following form:[2]

$$y_t = \beta E_t y_{t+1} - \sigma r_t$$

$$\pi_t = \beta E_t \pi_{t+1} + \gamma x_t,$$

where y, r, π, and x represent the logarithm of output, the real rate of interest, inflation, and a driving variable for inflation, respectively. The discount rate is β, and γ and σ are positive parameters.

When the output gap is taken as the variable driving inflation, and rational expectations are assumed, these two equations imply no inertia whatsoever in output or inflation beyond what is induced by the nominal rate of interest.[3] Both variables behave like stock prices: they jump in response to shocks.[4] The long and variable lags that seem to characterize links from monetary policy to output and inflation in the data are nowhere to be found in the models, and any inertial behavior in the economy can derive only from a perplexingly inertial monetary authority.

Of course, models with these characteristics have little hope of matching the key dynamic features of U.S. macroeconomic history, as a number of authors have now demonstrated.[5] It is not possible to remedy this defi-

[2] The output equation is identical to the Euler equation for consumption in a standard life-cycle/permanent income model with a time-varying real interest rate, as the capital stock is assumed to be fixed.

[3] The same conclusion will be true with marginal cost as the driving variable for inflation, if marginal cost is in turn a purely forward-looking function of state variables.

[4] Note that the equations for output and inflation are isomorphic to dividend-pricing models of stock prices.

[5] See, for example, Cogley and Nason (1995), Nelson (1998), Estrella and Fuhrer (2002), and Fuhrer (2000).

ciency by substituting real marginal cost for the output gap in the inflation equation.[6] If marginal cost is the proper determinant of inflation, and it in turn is determined by a purely forward-looking relation, the resulting model has exactly the same undesirable properties.

The problem is that the data show considerably more inertia, in both prices and quantities, than the models imply. Some authors have proposed ad hoc solutions to this problem. For example, some have suggested that a fraction of price setters or consumers exhibit "rule-of-thumb" or "backward-looking" behavior.[7] But this solution is a capitulation, as it explains all of the persistence in inflation or consumption changes with machinery that is not derived from the assumed objectives of consumers or firms.[8] The more important these rule-of-thumbers are, the less structure there is to the structural model into which they are inserted. The more rule-of-thumbers there are, the less representative is the utility function of the optimizing consumers, and the less hope we have of deriving optimal policy from a well-defined social welfare function, because rule-of-thumbers do not maximize anything but simply follow the rule. Who knows which policies they would prefer?

A host of other authors have turned to habit formation to solve some of the dynamic models' problems.[9] This approach seems more promising, but it should be of considerable concern that studies of micro consumption data do not find evidence of habit formation.[10] This raises the possibility that habit formation in aggregate consumption stands for some other underlying consumer behavior yet to be identified. This, too, should raise concerns about the welfare implications derived from such models.

Many central banks augment models based on pure theory with lags, either by assuming a fraction of backward-looking agents or by positing high-order adjustment costs in most sectors of the economy. These augmentations are empirically driven and represent in my view a somewhat uncomfortable balance between existing theories and the knowledge that

[6] As in Galí and Gertler (1999) and Galí, Gertler, and Lopez-Salido (2001).

[7] Galí and Gertler (1999) and Galí, Gertler, and Lopez-Salido (2001) invoke such behavior on the part of price setters, and Amato and Laubach (2002) on the part of consumers.

[8] The "rational" agents in these economies, of course, take account of the inertial behavior of their less rational counterparts, so this statement requires a minor qualification.

[9] On habit formation in asset markets see Campbell and Cochrane (1999); on habit formation and economic growth see Carroll, Overland, and Weil (2000); on habit formation and monetary policy see Fuhrer (2000).

[10] Dynan (2000).

the data exhibit significant persistence that is inconsistent with the theories. If most of the persistence in the model derives from nonoptimizing, non-theory-based mechanisms, what are we to make of the welfare conclusions from these models?

Perhaps as disturbing as the shortage of theoretically sound, empirically verified models is the fact that the sources of inflation distortions in aggregate models remain rather murky.[11] Within one tradition there is little agreement on why money should exist, and therefore little agreement on how inflation acts as a tax on it or distorts its use in one activity or another. Does money enter the utility function? Is it required in advance for certain transactions? Does it help us to economize on shopping time? Another strand of the literature asserts that the price dispersion caused by the pricing decisions of imperfectly competitive firms is the source of the inflation distortion. This story, however, derives from an assumed fixity in firm pricing behavior that seems hard to justify. In yet another strand, the modeler simply assumes that the central bank has a distaste for inflation, even though inflation causes little or no explicit distortion in the model. This is the case in the Federal Reserve's FRB/US model. Finally, in another tradition, the one that I align myself with, all these ways of modeling inflation distortions are viewed with considerable suspicion.

I would conclude, first, that the models do not fit the data, and second, that they do not provide plausible descriptions of the welfare losses that policy is trying to minimize. The two conclusions are related, of course, in that models that fit so poorly could not really be good reflections of the underlying objectives of consumers and firms.

Where does this leave us? In my view, we have a few significant challenges to tackle over the coming decades:

—We should keep the empirical standard for model development quite high. Estimated Euler equations tell us little about the dynamic properties of the macroeconomic variables we are trying to understand. We cannot claim to have identified macroeconomic relationships unless we have fully tested their dynamic implications, and that can really be done only with system methods.

[11] We do have stories that we tell undergraduates—about nominal contracts, downward nominal wage rigidity, and the zero bound on nominal interest rates—but for the most part, whether these stories make rigorous sense or not, they are not reflected in our aggregate models.

—We need to continue the program of research, begun by Robert Lucas and others, that seeks to understand the microeconomic underpinnings of macroeconomic regularities. But we need to broaden the scope of the inquiry beyond small wrinkles in the canonical utility functions and constraints. Dare I suggest that behavioral economists, or even psychologists, may possess some wisdom about economic decisionmaking from which we could benefit? I believe it is time to reexamine the commonly accepted assumptions regarding utility maximization, profit maximization, and formation of expectations that we routinely build into our macroeconomic models, over and over again.

—Let's not be in such a rush. Economists seem to be in a hurry to compute optimal policy results from a microfounded model, without first being sure that the model fits the data, has identified the underlying behaviors we are looking for, and embodies plausible descriptions of household and firm objectives.

In this context Sims' analysis of the role of models in monetary policy deliberations in the United States and around the world is useful and interesting. As always, I learned a lot from reading his paper. I agree with much of what he says, but I have questions about a few of his conclusions.

Sims replicates the finding of Romer and Romer that the Federal Reserve's inflation forecasts are superior to private forecasts,[12] and he adds to that analysis by pointing out that "information content" regressions may give a misleading indication of relative forecast performance. His analysis of possible reasons for the Federal Reserve's forecasting edge is a potentially valuable addition to the work of Romer and Romer. But, as Sims himself suggests, this analysis is not conclusive, and thus several questions remain.

First, why is it that the Federal Reserve's inflation forecasts are so much better than others, but their output forecasts are not?[13] This is puzzling, because the Federal Reserve relies on forecasting frameworks, both model-based and judgmental, that link real "gap" variables (unemployment or output) to inflation. The medium-term direction of inflation in these models should largely depend on the current state of the real gap variables. If, on average, the Federal Reserve's forecasts of these vari-

[12] Romer and Romer (2000).

[13] This fact is established in Romer and Romer's (2000) paper but might change if Sims' measure were computed.

ables are no better than those of the rest of the profession, whence comes its inflation forecasting prowess?

Second, Sims suggests that the Federal Reserve staff's intensive examination of "special factors" affords them some short-run advantage. To be sure, the staff devotes significant resources to analyzing the impact of impending tax, methodological, and relative price changes on aggregate price indices. But the question is how many of these factors have persistent effects on inflation. A good number likely have persistent effects only on the price level, with transient effects on the inflation rate, and thus should not improve longer-term forecasts. Others, such as the large swings in oil prices in the 1970s and early 1980s, likely did have persistent effects on inflation. To the extent that the Federal Reserve staff had a more accurate view of the effects of oil price developments on inflation, this could have resulted in superior forecast performance. Thus an important unresolved question is to what extent the expertise of Federal Reserve specialists results in better understanding of transient or persistent influences on inflation in this sample.

Third, Sims considers the possibility that the Federal Reserve's inflation forecasting edge arises from its superior knowledge of the likely course of its own future policy actions. But regardless of what the staff knows, the Green Book forecast is not a true "forecast," but rather a projection conditioned on an assumed path for the federal funds rate. And in many if not most of the cases in Sims' sample, the funds rate path is constructed precisely to show what would happen if the funds rate were unchanged, even when that may not be the prevailing expectation in private markets or inside the Federal Reserve itself. This is not to say that the assumed path never reflects the staff's best forecast of future policy. But for the many projections that do not condition on the best funds rate forecast, this complicates Sims' inferences about the sources of forecast performance.

Finally, on Sims' methodological points, it is surely true that the policy process necessarily involves a mixture of judgmental and model inputs, confronts uncertainty both in constructing analytical frameworks and in implementing policy, and will thus at some stage be ripe for Bayesian econometric techniques. But many of the advantages of the Bayesian techniques that Sims cites apply primarily to the large-model, large-staff processes currently in place at central banks. My sense is that critical advances in central bank models and understanding are more likely to be found through work on smaller structural models that improve on the primary behavioral relationships at work in the models, as suggested above. For these models, which impose fairly strict theoretical restrictions, the problems of instrument reduction and combining subjective judgment with econometric models do not loom as large. In my judgment, relatively straightforward classical econometric techniques would have, and indeed have, allowed us to distinguish between successful and unsuccessful economic models of this type.[14]

Thus the econometric methods that Sims advocates in the paper might be viewed as refinements of a model that already has cleared the hurdles of incorporating well-identified and data-consistent microfoundations. As a result, it may be too early in the development of monetary policy models to make best use of these Bayesian techniques. When monetary policy models have matured substantially, I would imagine that Sims' suggestions will be valuable enhancements to model design and implementation at central banks.

General discussion: Several panelists commented on the way models are used in the policymaking process. Benjamin Friedman agreed with Sims that the small ECB models focusing on money were not the rationale for the second pillar of the ECB, but rather that the ECB wanted a second pillar for independent reasons. The models were built in order to support the second pillar. In Willem Buiter's view the only reliable parts of the Bank of England's model were the accounting identities. However, he thought the model still served a useful role as a framework for discussions and in guaranteeing a minimum level of consistency. The model forced the monetary policy committee and staff to express their views in a coherent way—one could not forecast output growth of 4 percent along with a 3 percent increase in hours worked and 5 percent productivity growth. Although the models did assist in making intelligent guesses about the future course of the economy, they did not play a systematic role in preparing forecasts. Indeed, at the Bank of England the model was adjusted to fit what the committee and staff thought was a reasonable projection, even to the extent of eliminating whole equations. In the same

[14] Of course, although the specific Bayesian techniques mentioned in the paper may not be of relevance for small, highly restricted structural models, a Bayesian approach might well be.

vein, David Reifschneider stressed that, at the Federal Reserve, large-scale models were not primarily used for forecasting but for facilitating discussions among policymakers. He believed the model simulations showing different ways in which the economy might unfold were very useful. He noted that although the staff sometimes attach probabilities to these scenarios, policymakers alone decide how they use this information.

William Nordhaus observed that although the four central banks Sims examined have very different cultures, all rely on teams of sectoral experts. Results of some studies suggested that, in the United States, the experts outperformed the large models, and Nordhaus conjectured that this largely reflected their ability to react faster to crises and structural changes. One episode in which this appeared to have happened was the fall in oil prices in 1986. Experts' use of simple sectoral models allows a high degree of flexibility. These models can be easily adjusted to new situations, whereas much more effort is needed to adapt complex simultaneous-equations models in a similar way. Hence decentralization and specialization are both justified and likely to improve performance. Nordhaus noted that in other areas, such as climate and environmental modeling, delegation of this kind is common. Ben Bernanke agreed with the view that experts have a tremendous amount of information at the microeconomic or sectoral level and that information aggregation is central to successful forecasting. So, for short-term forecasting, a very limited amount of structure might indeed be optimal.

Olivier Blanchard agreed with Sims that Bayesian methods should be used more widely. But he wondered whether the key to solving problems with the existing models really lies in substituting Bayesian for non-Bayesian econometrics. He described four criticisms levied at the MPS model that were not related to econometric methodology. He thought most of these criticisms could be dealt with, and that the resulting model would be superior to the current U.S. and Canadian models. One of these criticisms was that the MPS model was prone to explosive behavior. Blanchard believed that a modest amount of additional structure would solve this problem, giving the model a balanced growth path while retaining the informational advantages of using experts for various sectors. A second criticism was that the original model's treatment of expectations was inadequate. That has been resolved, and we now know how to estimate equations explicitly incorporating expectations and how to do simulations with rational expectations. A third criticism was the imposition of so many zero restrictions in the equations when theory suggests including any variable that could affect expectations. Blanchard thought this problem had been oversold and that, after controlling for expectations, zero restrictions can reasonably be imposed in many variables. A fourth criticism, which Jeffrey Fuhrer had emphasized in his comment, is that the model's equations, particularly those relating to dynamics, lacked theoretical justification. Conversely, the dynamics delivered by theory often do not match the dynamics observed in the data. Blanchard saw no immediate way to avoid using ad hoc dynamics until more-adequate theory is developed. He agreed that doing so inhibits formal welfare evaluations, but he did not see this as a fatal flaw. Minimization of the variance of output is a legitimate objective of policy, and a model with an ad hoc but realistic specification of dynamics is a useful tool for that purpose.

Friedman supported Sims' view that the Lucas critique was not very important at the one- or two-year horizons that are of most concern to policymakers. He believed that within this time frame it is reasonable to assume that the public neither perceives nor adapts to a shift in policy regime. He thought it made little difference whether the monetary authority says "This is what we're going to do in this situation" or "We are henceforth going to have a rule that says this is what we are going to do every time we are in this situation."

Reifschneider observed that any discussion about changing the models should also include a cost-benefit analysis and a recognition of the limited resources available to central banks. He noted that staff at the Federal Reserve are aware of the many shortcomings of the models under discussion. But a great many people already work full time on the models, because model building and maintenance remain a complex task. To implement all the suggestions he had heard, the Federal Reserve's human resources devoted to this work would have to be increased dramatically. Other central banks, which typically have much smaller staffs than the Federal Reserve, are in an even more difficult position. Complicating this resource problem is a lack of support from the academic community, which has largely abandoned large-scale modeling. He hoped that this would change and that academic economists would turn to providing practical assistance to central banks in their modeling work.

References

Atkeson, Andrew, and Lee E. Ohanian. 2001. "Are Phillips Curves Useful for Forecasting Inflation?" *Federal Reserve Bank of Minneapolis Quarterly Review* 25(1): 2–11.

Amato, Jeffery D., and Thomas Laubach. 2002. "Rule-of-Thumb Behaviour and Monetary Policy." Finance and Economics Discussion Series 2002-5. Washington: Board of Governors of the Federal Reserve System.

Black, Robert, and others. 1994. "The Bank of Canada's New Quarterly Projection Model Part 1, The Steady-State Model: SSQPM." Technical Report 72. Ottawa: Bank of Canada.

Blix, Martin, and Peter Sellin. 2000. "A Bivariate Distribution for Inflation and Output Forecasts." Working Paper 102. Stockholm: Sveriges Riksbank.

Brock, Willard A., and Steven N. Durlauf. 2001. "Growth Empirics and Reality." *World Bank Economic Review* 15(2): 229–72.

Campbell, John Y., and John H. Cochrane. 1999. "By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior." *Journal of Political Economy* 107(2): 205–51.

Carroll, Christopher D., Jody R. Overland, and David N. Weil. 2000. "Saving and Growth with Habit Formation." *American Economic Review* 90(3): 341–55.

Cogley, Tim, and James M. Nason. 1995. "Output Dynamics in Real-Business-Cycle Models." *American Economic Review* 85(3): 492–511.

Coletti, Donald, and others. 1996. "The Bank of Canada's New Quarterly Projection Model Part 3, The Dynamic Model: QPM." Technical Report 75. Ottawa: Bank of Canada.

Dehejia, Rajeev. Forthcoming. "Program Evaluation as a Decision Problem." *Journal of Econometrics.*

Doppelhofer, Gernot, Ronald I. Miller, and Xavier Sala-i-Martin. 2000. "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach." Working Paper 7750. Cambridge, Mass.: National Bureau of Economic Research.

Draper, David. 1995. "Assessment and Propagation of Model Uncertainty." *Journal of the Royal Statistical Society* (series B) 57: 45–97.

Dynan, Karen E. 2000. "Habit Formation in Consumer Preferences: Evidence from Panel Data." *American Economic Review* 90(3): 391–406.

Estrella, Arturo, and Jeffrey C. Fuhrer. 2002. "Dynamic Inconsistencies: Counterfactual Implications of a Class of Rational Expectations Models." *American Economic Review* 92(4): 1013–28.

Fagan, Gabriel, Jerome Henry, and Ricardo Mestre. 2001. "An Area-Wide Model (AWM) for the Euro Area." Working Paper 42. Frankfurt: European Central Bank.

Fair, Ray C., and Robert C. Shiller. 1989. "The Informational Content of Ex Ante Forecasts." *Review of Economics and Statistics* 71(2): 325–31.

Fernandez, Carmen, Eduardo Ley, and Mark F. I. Steel. 2001a. “Benchmark Priors for Bayesian Model Averaging.” *Journal of Econometrics* 100(2): 381–427.

———. 2001b. “Model Uncertainty in Cross-Country Growth Regressions.” *Journal of Applied Econometrics* 16(5): 563–76.

Freedman, D. 1991. “A Rejoinder to Berk, Blalock, and Mason.” *Sociological Methodology* 21: 353–58.

Fuhrer, Jeffrey C. 2000. “Habit Formation in Consumption and Its Implications for Monetary Policy Models.” *American Economic Review* 90(3): 367–90.

Galí, Jordi, and Mark Gertler. 1999. “Inflation Dynamics: A Structural Econometric Analysis.” *Journal of Monetary Economics* 44(2): 195–222.

Galí, Jordi, Mark Gertler, and David J. Lopez-Salido. 2001. “European Inflation Dynamics.” *European Economic Review* 45(7): 1237–70.

Geraats, Petra M. 2001. “Why Adopt Transparency? The Publication of Central Bank Forecasts.” Working Paper 41. Frankfurt: European Central Bank.

Haavelmo, Trygve. 1944. “The Probability Approach in Econometrics.” *Econometrica* 12(suppl.): 1–115.

Hurwicz, Leonid. 1962. “On the Structural Form of Interdependent Systems.” In *Logic, Methodology and Philosophy of Science*. Stanford University Press.

Kiley, Michael T. 2001. “Business Investment in the Federal Reserve Board’s U.S. Model (FRB/US): Specification and Implications.” Unpublished paper. Washington: Board of Governors of the Federal Reserve.

Manski, Charles F. 2000. “Identification Problems and Decisions under Ambiguity: Empirical Analysis of Treatment Response and Normative Analysis of Treatment Choice.” *Journal of Econometrics* 95(2): 415–42.

Nelson, Edward. 1998. “Sluggish Inflation and Optimizing Models of the Business Cycle.” *Journal of Monetary Economics* 42(2): 303–22.

Phillips, Peter C. B. 1991. “To Criticize the Critics: An Objective Bayesian Analysis of Stochastic Trends.” *Journal of Applied Econometrics* 6: 333–64.

———. 1996. “Econometric Model Determination.” *Econometrica* 64(4): 763–812.

Poloz, Steven S., David Rose, and Robert J. Tetlow. 1994. “The Bank of Canada’s New Quarterly Projection Model (QPM): An Introduction.” *Bank of Canada Review*, pp. 23–38.

Quinn, Meghan. 2000. *Economic Models at the Bank of England: September 2000 Update*. London: Bank of England.

Raftery, Adrian E., David Madigan, and Jennifer A. Hoeting. 1997. “Bayesian Model Averaging for Linear Regression Models.” *Journal of the American Statistical Association* 92(437): 179–91.

Reifschneider, David L., David J. Stockton, and David W. Wilcox. 1997. “Econometric Models and the Monetary Policy Process.” *Carnegie-Rochester Conference Series on Public Policy* 47: 1–37.

Romer, Christine D., and David H. Romer. 2000. "Federal Reserve Information and the Behavior of Interest Rates." *American Economic Review* 90(3): 429–57.

Sargent, Thomas J. 1984. "Autoregressions, Expectations, and Advice." *American Economic Review Papers and Proceedings* 74: 408–15.

Sims, Christopher A. 1987. "A Rational Expectations Framework for Short Run Policy Analysis." In *New Approaches to Monetary Economics*, edited by William Barnett and Kenneth Singleton. Cambridge University Press.

———. 1989. "Modeling Trends." In *Proceedings of the American Statistical Association Annual Meetings.* eco072399b.princeton.edu/yftp/trends/asa889.pdf.

———. 1993. "A Nine-Variable Probabilistic Macroeconomic Forecasting Model." In *Business Cycles, Indicators, and Forecasting*, edited by James H. Stock and Mark W. Watson. University of Chicago Press.

———. 1996. "Bayesian Inference for Multivariate Time Series with Trend." Discussion paper presented at the 1992 American Statistical Association Meetings. eco072399b.princeton.edu/yftp/trends/asapaper.pdf.

———. 2000. "Comments on Papers by Jordi Galí and by Stefania Albanesi, V. V. Chari, and Lawrence J. Christiano." Presented at the 8th World Congress of the Econometric Society, Seattle, August.

Sims, Christopher A., and Tao Zha. 1998. "Bayesian Methods for Dynamic Multivariate Models." *International Economic Review* 39(4): 949–68.

Smets, Frank, and Raf Wouter. 2002. "An Estimated Dynamic Stochastic General Equilibrium Model of the Euro Area." Working paper. Frankfurt: European Central Bank and Brussels: National Bank of Belgium.

Tinbergen, Jan. 1939. *Business Cycles in the United States of America, 1919–1932,* vol. 2 of *Statistical Testing of Business Cycle Theories.* Geneva: League of Nations.

Wald, Abraham. 1950. *Statistical Decision Functions.* New York: John Wiley.