

# **EVALUATING PROBABILITY FORECASTS FOR GDP DECLINES**

by

Kajal Lahiri  
Univeersity at Albany – SUNY

and

J. George Wang  
College of Staten Island – CUNY

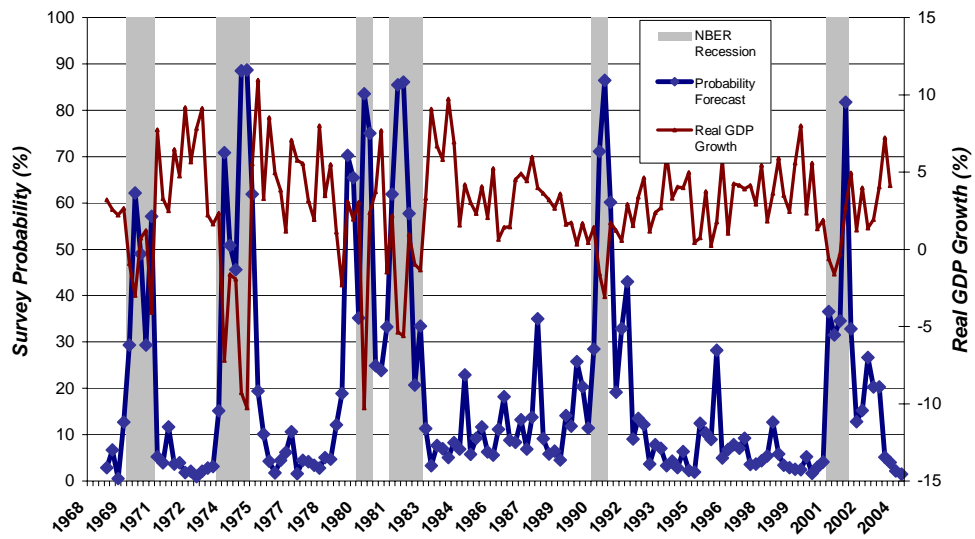
For presentation at the **FIRST MACROECONOMIC FORECASTING  
CONFERENCE (IFO-INSEE-ISAE): Rome, March 27, 2009.**

**Main Results:** Using evaluation methodologies for rare events from meteorology and psychology, we examine the value of probability forecasts of real GDP declines during the current and each of the next four quarters using data from the *Survey of Professional Forecasters*. We study the quality of these probability forecasts in terms of calibration, resolution, odds ratio, the relative operating characteristic (ROC), and alternative variance decompositions. Even though QPS and the calibration tests for perfect forecast validity for all five horizons were accepted, the other approaches clearly identify the longer-term forecasts (Q3-Q4) having no skill. For a given hit rate of (say) 90%, the associated high false alarm rates that underlie the longer-term forecasts make these unusable in practice. We find conclusive evidence that the shorter-term forecasts (Q0-Q2) possess significant skill in terms of all measures considered, even though they are characterized by excess variability.

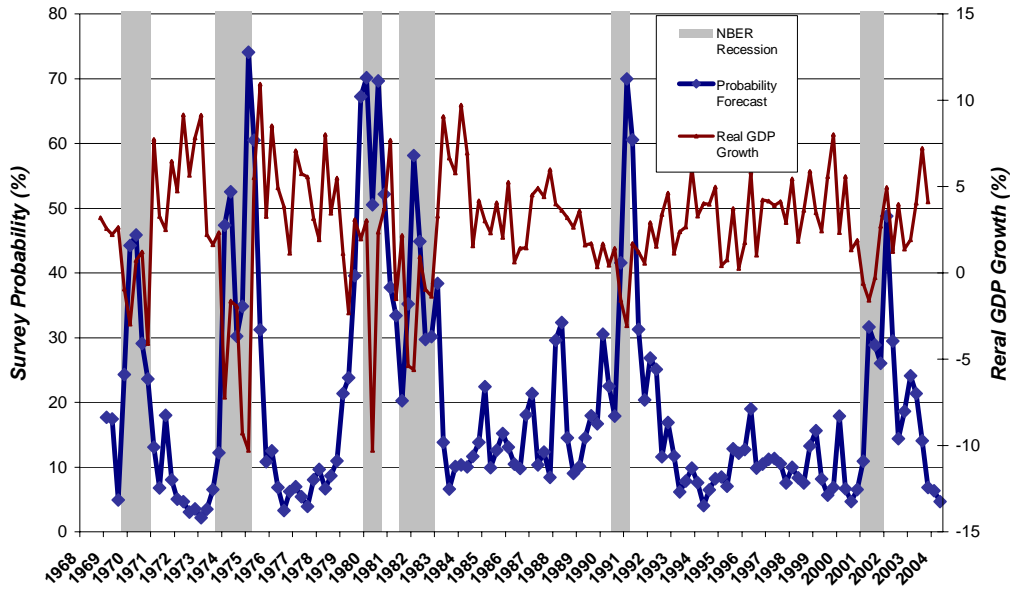
## **SPF PROBABILITY FORECASTS OF REAL GDP DECLINE**

The SPF probabilities for real GDP declines during the current and next four quarters are depicted against the real time real GDP growth in Figures 1a - 1e. The shaded bars represent the NBER defined recessions.

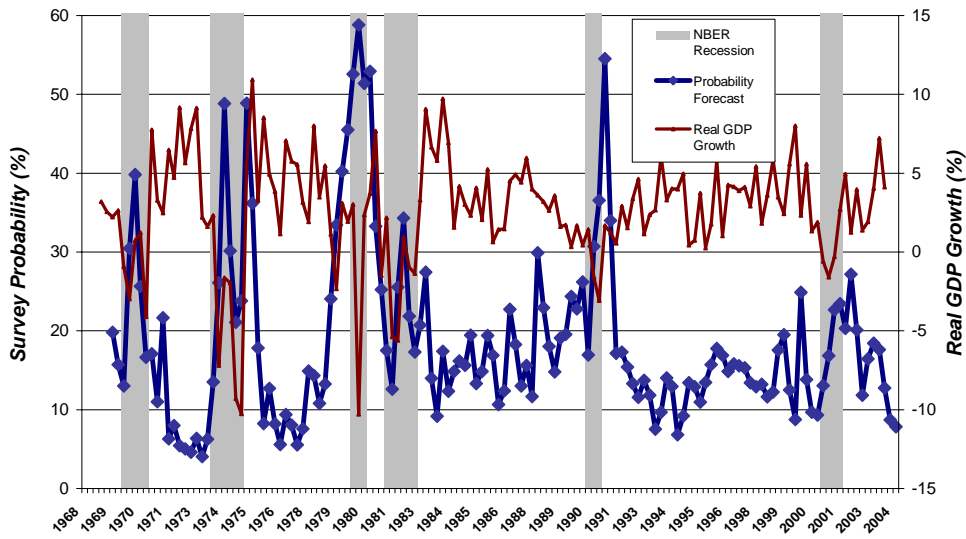
Fig. 1a: Probability of Decline in Real GDP in the Current Quarter



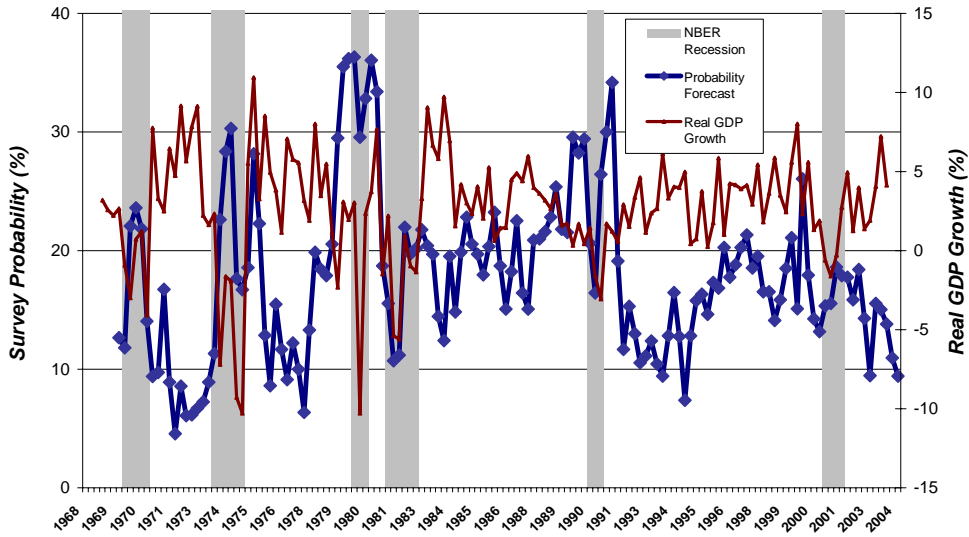
**Fig. 1b: Probability of Decline in Real GDP in the Following Quarter**



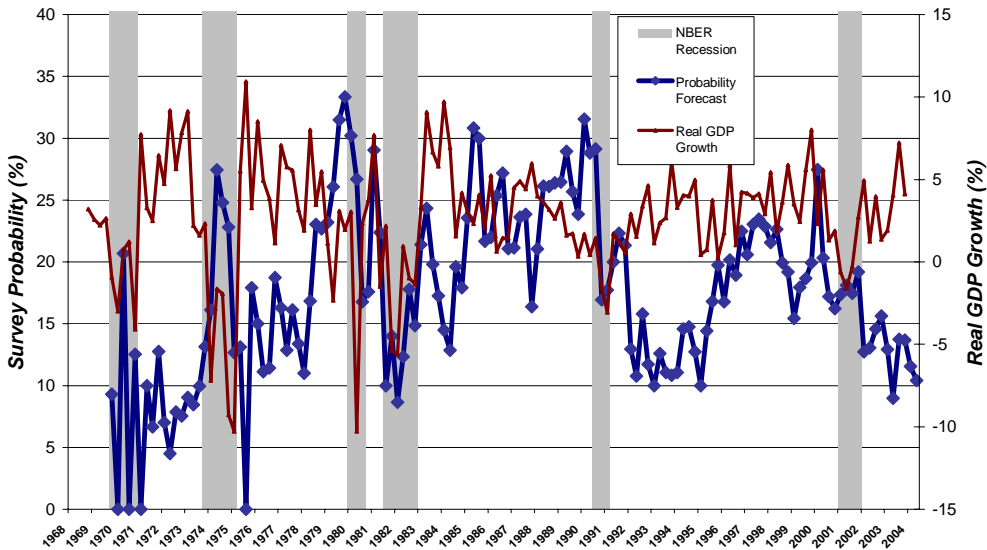
**Fig. 1c: Probability of Decline in Real GDP in Following Second Quarter**



**Fig. 1d: Probability of Decline in Real GDP in Following Third Quarter**



**Fig. 1e: Probability of Decline in Real GDP in Following Fourth Quarter**



From Figures 1a-1e, several notable patterns can be observed. First, the mean probabilities generated by the professional forecasters fluctuate over

time, varying from as high as 80% to as low as less than 5%. Secondly, high end of the mean probability tends to decrease as the forecasting horizon increases. As shown in the figures, the high-end probability decreases steadily from about 80% for the current quarter to only about 30% for three and four-quarter-ahead forecasts. All these observations suggest that the information content, hence the value, of the SPF probability forecasts may be horizon-dependent.

## **CALIBRATION OF *SPF* PROBABILITY FORECASTS**

### **Brier's Quadratic Probability Score**

A most commonly used measure is Brier's Quadratic Probability Score (*QPS*), a probability analog of mean squared error, *i.e.*:

$$QPS = 1/T \sum_{t=1}^T (f_t - x_t)^2 \quad (1)$$

where  $f_t$  is the forecast probability made at time  $t$ ,  $x_t$  is the realization of the event (1 if the event occurs and 0 otherwise) at time  $t$ .  $T$  is the total number of the observations or forecasting quarters in our case.

The *QPS* ranges from 0 to 1 with a score of 0 corresponding to perfect accuracy, and is a function only of the difference between the assessed probabilities and realizations. The calculated *QPS* for each forecasting horizon from the current quarter (Q0) to the next four quarters (Q1, Q2, Q3, and Q4) are calculated to be 0.077, 0.098, 0.103, 0.124, and 0.127, respectively.

### **Prequential Test for Calibration**

Dawid (1984) and Seillier-Moiseiwitsch and Dawid (1993) (henceforward SM-D) suggested a test for calibration-in-the-small when a sequence of  $T$  probability forecasts is grouped in probability intervals, *e.g.*, as in Tables 1 & 2.

**Table 1: Calculations for the Calibration Test: Quarter 0**

<b>Probability Interval</b>	<b>Midpoint</b>	<b>Frequency</b>	<b>Occurrence</b>	<b>Relative Frequency</b>	<b>Expectation</b>	<b>Weight</b>	<b>Test Statistic</b>	<b>Chi Square</b>
	$f_j$	$T_j$	$r_j$	$r_j/T_j$	$e_j = f_j T_j$	$w_j = T_j f_j (1 - f_j)$	$Z_j = (r_j - e_j) / \sqrt{w_j}$	$Z_j^2$
0.00 - 0.049	0.025	42	0	0.00	1.05	1.02	-1.04	1.08
0.05 - 0.149	0.100	51	0	0.00	5.10	4.59	-2.38	5.67
0.15 - 0.249	0.200	25	7	0.28	5.00	4.00	1.00	1.00
0.25 - 0.349	0.300	4	1	0.25	1.20	0.84	-0.22	0.05
0.35 - 0.449	0.400	3	2	0.67	1.20	0.72	0.94	0.89
0.45 - 0.549	0.500	6	2	0.33	3.00	1.50	-0.82	0.67
0.55 - 0.649	0.600	4	2	0.50	2.40	0.96	-0.41	0.17
0.65 - 0.749	0.700	3	1	0.33	2.10	0.63	-1.39	1.92
0.75 - 0.849	0.800	5	5	1.00	4.00	0.80	1.12	1.25
0.85 - 0.949	0.900	0	0	0.00	0.00	0.00	0.00	0.00
0.95 - 1.000	0.975	0	0	0.00	0.00	0.00	0.00	0.00
		<b>143</b>	<b>20</b>				$\chi^2 = \sum Z_j^2$	<b>12.68</b>

**Table 2: Calibration Tests for Q0-Q4**

<b>Midpoint</b>	$Z_j(0)$	$Z_j(1)$	$Z_j(2)$	$Z_j(3)$	$Z_j(4)$
<b>0.025</b>	-1.04	-0.55	-0.23	-0.16	-0.16
<b>0.1</b>	-2.38	-2.37	-1.45	-0.43	0.37
<b>0.2</b>	1.00	-0.92	-1.00	-1.10	-0.93
<b>0.3</b>	-0.22	1.01	1.41	0.28	-1.57
<b>0.4</b>	0.94	0.15	-0.61	-1.63	0.00
<b>0.5</b>	-0.82	0.00	-1.89	0.00	0.00
<b>0.6</b>	-0.41	-0.94	0.82	0.00	0.00
<b>0.7</b>	-1.39	-1.46	0.00	0.00	0.00
<b>0.8</b>	1.12	0.00	0.00	0.00	0.00
<b>0.9</b>	0.00	0.00	0.00	0.00	0.00
<b>0.975</b>	0.00	0.00	0.00	0.00	0.00
$\chi^2$	12.7	10.82	9.74	4.17	3.50
<b>QPS Test (N(0,1))</b>	-1.597	-1.481	-1.393	-1.137	-0.916

Given  $QPS = 1/T \sum_{t=1}^T (f_t - x_t)^2$ , SM-D showed how their calibration test could be converted to a test of  $QPS$  being significantly different from its expected value  $1/T \sum_{t=1}^T f_t(1-f_t)$  under the hypothesis of perfect forecast validity using a standard  $N(0,1)$  approximation for the distribution of

$$Y_n = [\sum_{t=1}^T (1-2f_t)(x_t - f_t)] / [\sum_{t=1}^T (1-2f_t)^2 f_t(1-f_t)]^{1/2} \quad (2)$$

When probability forecasts are grouped,  $Y_n$  can be calculated as:

$$Y_n = [\sum_{j=1}^{11} (1-2f_j)(r_j - e_j)] / [\sum_{j=1}^{11} T_j f_j(1-f_j)(1-2f_j)^2]^{1/2}. \quad (3)$$

The test results are reported in Table 2 as well. We find that, for all forecast horizons, none of the calculated statistics fall in the (one-sided) rejection region at the 5% significance level, which is consistent with the SM-D calibration test results that forecasts for all horizons satisfy the hypothesis of perfect forecast validity.

## **FURTHER DIAGNOSTIC VERIFICATIONS**

Some of the results from the calibration tests in the previous section may seem counter-intuitive. While the probability forecasts for the longer forecasting horizons, especially Q3 and Q4, never exceed 0.40 even when the event has already occurred, the SM-D test showed that they are well calibrated. This observation leads to a question of whether the calibration is an adequate measure of forecast validity, and why, if it is not. The issue may be analyzed using some alternative approaches.

## The Murphy Decomposition

In addition to calibration, there are several other features that also characterize good probability forecasts. Murphy (1972) decomposed the  $QPS$  or the Brier Score into three components. Rewriting  $QPS$  for grouped data

$$QPS(f, x) = (1/T) \sum_{j=1}^J \sum_{t=1}^{T_j} (f_j - x_{jt})^2, \quad (7)$$

the Murphy decomposition can be expressed as

$$QPS(f, x) = \bar{x}(1 - \bar{x}) + (1/T) \sum_{j=1}^J T_j (\bar{x}_j - f_j)^2 - (1/T) \sum_{j=1}^J T_j (\bar{x}_j - \bar{x})^2 \quad (8)$$

where  $\bar{x}_j = (1/T_j) \sum_{t=1}^{T_j} x_{jt}$  is the relative frequency of event's occurrence over  $T_j$

occasions with forecast  $f_j$ , *i.e.*,  $\bar{x}_j (= r_j/T_j)$  is an estimate of  $\mu_{x|f}$  using grouped data.

The first term on the RHS of (8) is the variance of the observations, and can be interpreted as the  $QPS$  of constant forecasts equal to the base rate. It represents forecast difficulty. The second term on the RHS of (8) represents the calibration or reliability of the forecasts. It can be interpreted as a labeling skill that expresses uncertainty. The third term on the RHS of (8) is a measure of resolution or discrimination; it refers to the ability of a set of probability forecasts to sort individual outcomes into probability groups which differ from the long-run relative frequency. A simple example explains the distinction between calibration and resolution. Suppose the event ( $=1$ ) occurs in every alternative period as 0, 1, 0, 1, .... Consider three sets of forecasts:  $F_1$  assigns 0.5 always;  $F_2$  assigns 0, 1, 0, 1, ..., and  $F_3$  assigns 1, 0, 1, 0,.... Here  $F_1$  and  $F_2$  are well calibrated, but  $F_2$  is perfect

whereas  $F_1$  is almost useless. Both  $F_2$  and  $F_3$  are perfectly resolved, but  $F_3$  is not well calibrated.  $F_3$  is more useful than  $F_1$  once we know how to calibrate  $F_3$ .

**Table 4: Murphy Decomposition**

<b>Lead Time</b>	<b>QPS (Accuracy)</b>	<b>= Uncertainty</b>	<b>+ Reliability</b>	<b>- Resolution</b>
<b>Q0</b>	0.0793	0.1211	0.0153	0.0572
<b>Q1</b>	0.1018	0.1219	0.0108	0.0308
<b>Q2</b>	0.1150	0.1226	0.0135	0.0210
<b>Q3</b>	0.1226	0.1233	0.0062	0.0069
<b>Q4</b>	0.1270	0.1218	0.0155	0.0103

Numerical values of the Murphy decomposition are given in Table 4 where we find that *QPS* improves by about 35%, 16% and 6% for the current (Q0), one quarter- (Q1), and 2-quarter-ahead (Q2) forecasts, respectively, over the constant relative frequency forecast (CRFF). The 3-quarter-ahead (Q3) forecasts are even with CRFF, and the *QPS* of the 4-quarter-ahead (Q4) forecasts are worse by nearly 4%.

The major contributor for the improvement in *QPS* is resolution, which helps to reduce the baseline *QPS* (CRFF) by about 47%, 25%, 17%, 6%, and 8% for Q0 to Q4, respectively. On the other hand, the miscalibration increases *QPS* of CRFF by 12%, 9%, 11%, 5% and 13%, respectively – they are relatively small for all forecast horizons. The improvement due to resolution is greater than the deterioration due to miscalibration for the up to 2-quarter-ahead forecasts, and the situation is opposite for the 4-quarter-ahead forecasts. In the case of 3-quarter-ahead forecasts, resolution and miscalibration pretty much cancel each other out.

In Figures 3a – 3e, the graph is split into two conditional likelihood distributions given  $x = 1$  (GDP decline) and  $x = 0$  (no GDP decline).

Figure 3a: Likelihood Diagram (Q0)

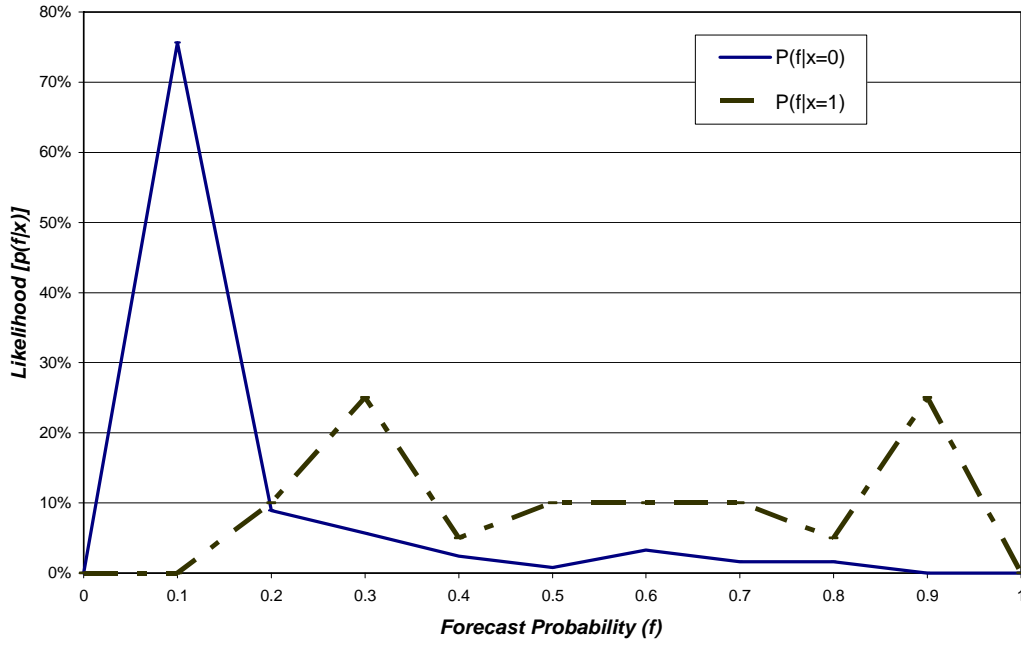


Figure 3b: Likelihood Diagram (Q1)

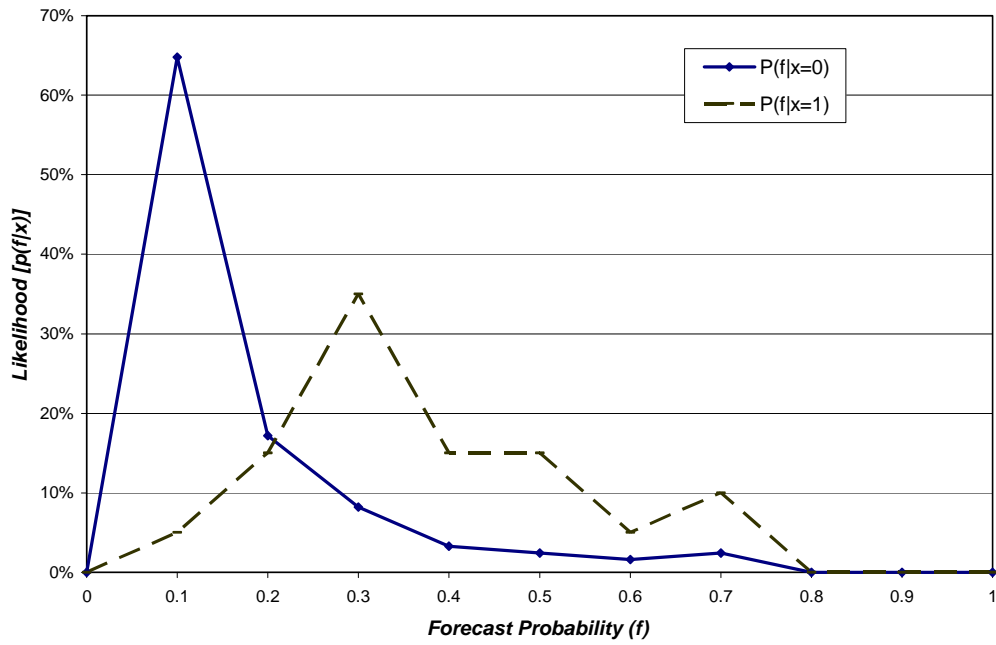


Figure 3c: Likelihood Diagram (Q2)

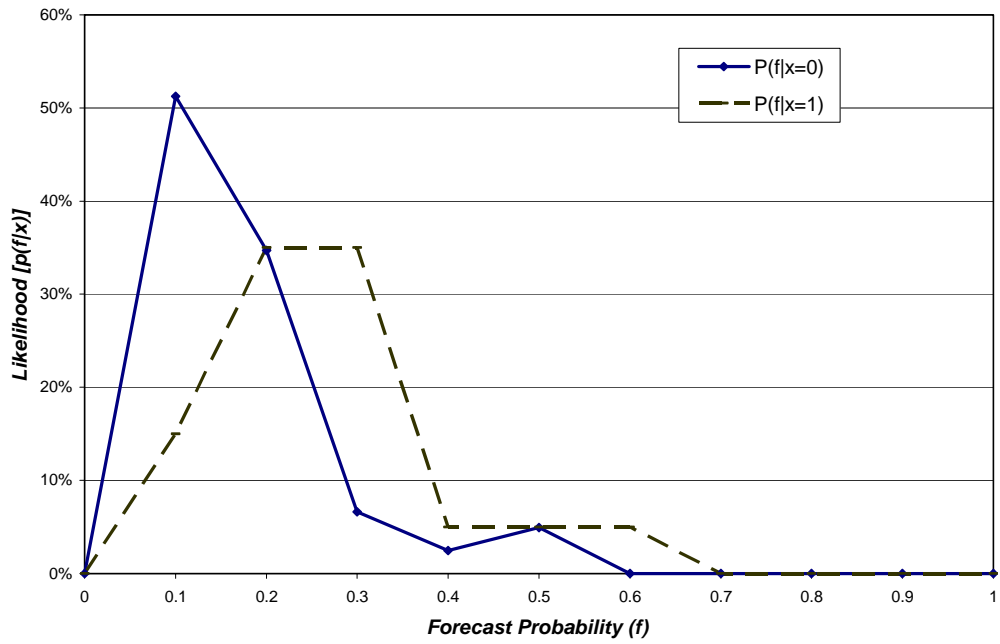


Figure 3d: Likelihood Diagram (Q3)

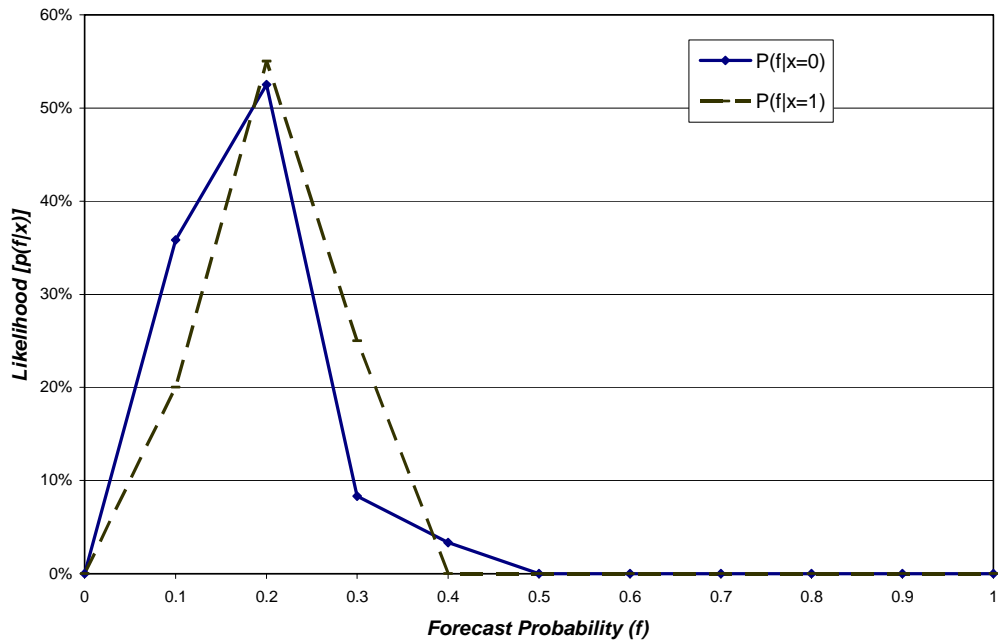
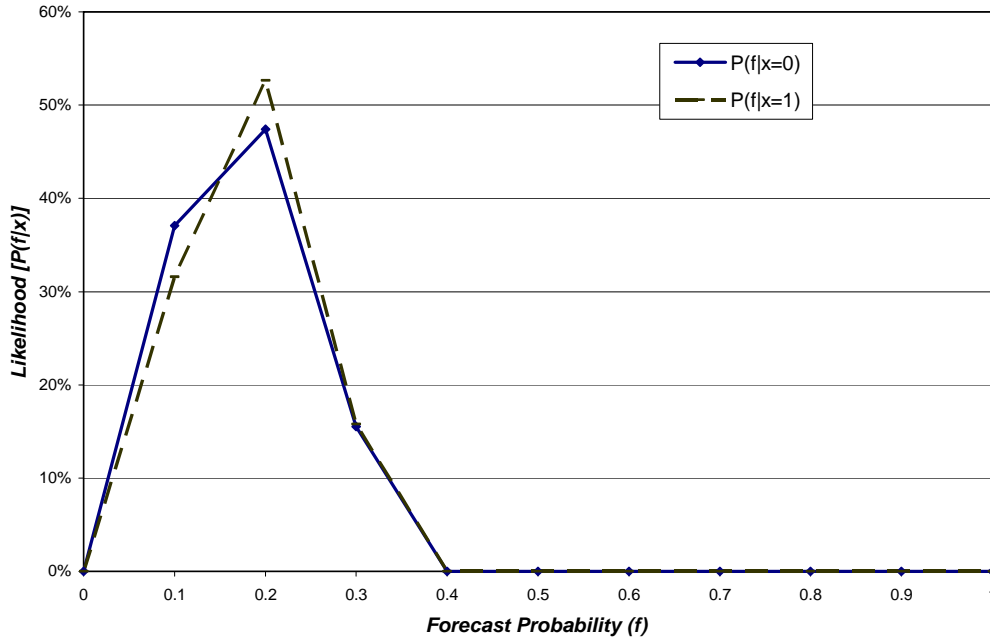


Figure 3e: Likelihood Diagram (Q4)



For these two conditional distributions, the means were calculated to be (0.56, 0.38, 0.26, 0.19 and 0.18) for  $x = 1$  and (0.14, 0.16, 0.17, 0.17 and 0.18) for  $x = 0$ , respectively. Good discriminatory forecasts will give two largely non-overlapping marginal distributions, and, in general, the ratio of their vertical differences should be as large as possible. While the shorter run forecasts (Q0-Q2) display better discriminatory power, the longer run forecasts (Q3-Q4) display poor discrimination due to the over-use of low probabilities during both regimes (*i.e.*,  $x = 0$  and  $x = 1$ ). So the two distributions overlap. In particular, the mean values for  $x = 1$  (GDP decline) and  $x = 0$  (no GDP decline) for the 4-quarter ahead forecasts (Q4) are almost identical.

### The Yates Decomposition

The calibration component in (8) can be written as:

$$(1/T) \sum_{j=1}^J T_j (f_j - \bar{x}_j)^2 = s_f^2 + (\bar{f} - \bar{x})^2 - 2s_{f\bar{x}} + (1/T) \sum_{j=1}^J T_j (\bar{x}_j - \bar{x})^2 \quad (9)$$

where  $\bar{f}$ ,  $s_f^2$  and  $s_{f\bar{x}}$  are the sample forecast mean, variance and covariance respectively. Since the last term in equation (9) is the resolution component in (8), Yates (1982) and Yates and Curley (1985) have argued that the calibration and resolution components in the Murphy decomposition are algebraically confounded with each other, and suggested a covariance decomposition of  $QPS$  that is more basic and more revealing than the Murphy decomposition, see also Björkman (1994) and Yates (1994). The so-called Yates decomposition is written as:

$$QPS(f, x) = \mu_x(1 - \mu_x) + \Delta\sigma_f^2 + \sigma_{f,\min}^2 + (\mu_f - \mu_x)^2 - 2\sigma_{f,x} \quad (10)$$

where  $\sigma_{f,\min}^2 = (\mu_{f|x=1} - \mu_{f|x=0})^2 \mu_x(1 - \mu_x)$ , and  $\Delta\sigma_f^2 = \sigma_f^2 - \sigma_{f,\min}^2$ .

As noted before, the outcome index variance  $\sigma_x^2 = \mu_x(1 - \mu_x)$  provides a benchmark reference for the interpretation of  $QPS$ . The conditional minimum forecast variance  $\sigma_{f,\min}^2$  reflects the double role that the variance of the forecast plays in forecasting performance. Even though minimization of  $\sigma_f^2$  will reduce  $QPS$ , this minimum value of forecast variance will be achieved only when a constant forecast is offered. But a constant forecast would lead to zero covariance between the forecast and event, which will, in turn, increase  $QPS$ . So the solution is to minimize the forecast variance given the covariance that demonstrates the fundamental forecast ability of the forecasters. The conditional minimum value of forecast variance (*i.e.*,  $\sigma_{f,\min}^2 = \sigma_f^2$ ) is achieved when the forecaster has perfect foresight such that he or she can exhibit perfect discrimination of the instances in which the event does and does not occur.

Since  $\Delta\sigma_f^2 = \sigma_f^2 - \sigma_{f,\min}^2$ , the term may be considered as the excess variability in forecasts. If the covariance indicates how responsive the forecaster is to information related to an event's occurrence,  $\Delta\sigma_f^2$  might reasonably be taken as a reflection of how responsive the forecaster is to information that is not related to the event's occurrence.

Using the SPF probability forecasts, the components of equation (10) were computed and presented in Table 5.

**Table 5: Yates Decomposition**

<b>Lead Time</b>	<b><math>QPS =</math></b>	<b><math>VAR(x) +</math></b>	<b><math>\Delta VAR(f) +</math></b>	<b><math>MinVAR(f) +</math></b>	<b><math>(\mu_f - \mu_x)^2 -</math></b>	<b><math>2 * COVAR(f, x)</math></b>
<b>Q0</b>	0.0769	0.1203	0.0330	0.0211	0.0033	0.1008
<b>Q1</b>	0.0977	0.1210	0.0212	0.0059	0.0032	0.0536
<b>Q2</b>	0.1146	0.1217	0.0113	0.0009	0.0020	0.0214
<b>Q3</b>	0.1227	0.1224	0.0046	0.0001	0.0012	0.0057
<b>Q4</b>	0.1265	0.1209	0.0040	0.0000	0.0016	0.0001

For the shorter forecasting horizons up to 2-quarters (Q0-Q2), the overall *QPS* values are less than the constant relative frequency forecast variance, which demonstrate the absolute skillfulness of the SPF probability forecasts. For the longer run forecasting horizons (Q3-Q4), the overall *QPS*s are slightly higher than those of the constant relative frequency forecast. The primary contributor of the performance is the covariance term that helps reduce the forecast variance by almost 84%, 44%, 18% and 5% for up to 3-quarter-ahead forecasts, but makes no contribution for the 4-quarter-ahead forecasts. The covariance reflects the forecaster's ability to make a distinction between individual occasions in which the event might or might not occur. It assesses the sensitivity of the forecaster to specific cues that are

indicative of what will happen in the future. It also shows whether the responsiveness to the cue is oriented in the proper direction. This decomposition is another way of reaching the same conclusion as the decomposition of skill score in Table 3a.

The excess variability of the forecasts,  $\Delta\sigma_f^2 = \sigma_f^2 - \sigma_{f,\min}^2$ , for each horizon is found to be 0.0330, 0.0212, 0.0113, 0.0046, and 0.0040, respectively. Compared to the overall forecast variances 0.0541, 0.0272, 0.0123, 0.0047, and 0.004, the excess variability's of SPF probability forecasts are 61%, 77%, 91%, 97% and 100% for Q0-Q4 forecasts, respectively. Thus, they are very high, and this means that the subjective probabilities are scattered unnecessarily around  $\mu_{f|x=1}$  and  $\mu_{f|x=0}$ . Since the difference in conditional means,  $\mu_{f|x=1} - \mu_{f|x=0}$ , are very close to zero for Q3-Q4 forecasts, all of their variability is attributed to excess variability. Assigning low probabilities in periods when GDP actually fell seems to be the root cause of the excess variance. The Yates decomposition gives this critical diagnostic information about the SPF probability forecasts that the Murphy decomposition could not. Overall, both the Murphy and Yates decompositions support the usefulness of shorter run SPF probabilities as predictors of negative real GDP growth, and suggest ways of improving the forecasts, particularly at short-run horizons.

### **RELATIVE (RECEIVER) OPERATING CHARACTERISTIC (ROC)**

The Murphy and Yates decompositions analyzed the structure of the total forecast error and the impact or the relative contribution of each component to the total error, but they could not provide a stand-alone single measure for

the discrimination ability of a forecast. In evaluating rare event probabilities, it is crucial to minimize the impact of the predominant outcome on the outcome score. More specifically, the impact of correctly identifying the frequent event, which is the primary source of the hedging, should be minimized. So a better approach to forecast performance should concentrate on the hit rate and false alarm rate of the infrequent event, instead of the “percentage correctly predicted” that is the very basis of *QPS*, cf. Doswell *et al* (1990) and Murphy (1991).

A simple and often-used measure of forecast skill, the Kuipers (or sometimes referred to as Pierce skill score) score (*KS*), is obtained by taking the difference between the hit rate (*H*) and the false alarm rate (*F*), where *H* is the proportion of times an event was forecast when it occurred, and *F* is the proportion of times the event was forecast when it did not occur.

**Table 6: Schematic Contingency Table**

<b>Event Forecasted</b>	<b>Event Observed</b>		<b>Total</b>
	<b>Occur</b>	<b>Not Occur</b>	
<b>Yes</b>	<b><i>a</i> (hit)</b>	<b><i>b</i> (false alarm)</b>	<b><i>a+b</i></b>
<b>No</b>	<b><i>c</i> (miss)</b>	<b><i>d</i> (correct rejection)</b>	<b><i>c+d</i></b>
<b>Total</b>	<b><i>a+c</i></b>	<b><i>b+d</i></b>	<b><i>a+b+c+d = T</i></b>

Given a decision threshold  $w$ , the contingency table for successes and failures for the event can be written as in Table 6. Then the Kuipers score can be calculated as  $H - F = (ad - bc) / ((a + c)(b + d))$ . Assuming independence of the hit and false alarm rates, the asymptotic standard error of the Kuipers

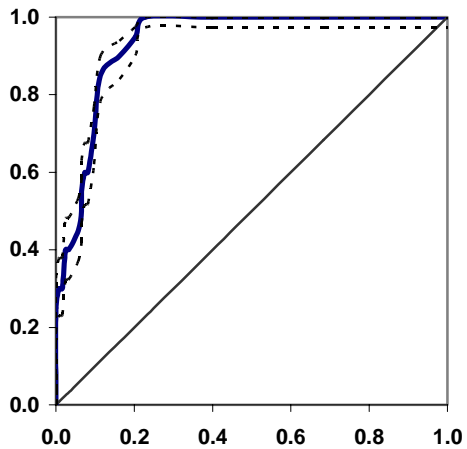
score is given by  $\sqrt{(H(1-H)/(a+c)) + ((F(1-F)/(b+d))}$ ; see Agresti (1996). Alternatively, based on the market-timing test of Pesaran and Timmermann (1992), Granger and Pesaran (2000) have suggested an alternative test for the significance of the Kuipers test,  $PT = \sqrt{T}KS / \sqrt{P_x(1-P_x)/\bar{x}(1-\bar{x})}$ , where  $P_x = \bar{x}H + (1-\bar{x})F$ . Stephenson (2000) notes that if one of the two elements in a column of the contingency table is very large (*e.g.*,  $d$ ), then Kuipers skill score effectively disregards the other element (*e.g.*,  $b$ ) almost completely. This can be a limitation of the Kuipers score in evaluating rare event forecasts.

Rather, the forecast skill can better be judged by comparing the odds of making a good forecast (a hit) to the odds of making a bad forecast (a false alarm), *i.e.*, by using the odds ratio  $\theta = \{H/(1-H)\}/\{F/(1-F)\}$  which is simply equal to the cross-product ratio  $(ad)/(bc)$  obtainable from the contingency table. The odds ratio is unity when the forecasts and the realizations are independent or  $KS=0$ , and can be easily tested for significance by considering the log odds that is approximately Normal with a standard error given by  $\sqrt{1/a+1/b+1/c+1/d}$ .  $KS$  and  $\theta$  are reported in Table 7 for relevant values of the decision threshold  $w$ .

One important but often overlooked issue in the evaluation of probability forecasts is the role of the selected threshold. The performance of a probability forecast in terms of discrimination ability is actually the result of the combination of the intrinsic discrimination ability of a forecasting system and the selection of the threshold. In these regards, Relative (or Receiver) Operating Characteristic (ROC) is a convenient descriptive approach, but unfortunately has drawn little attention in econometrics.

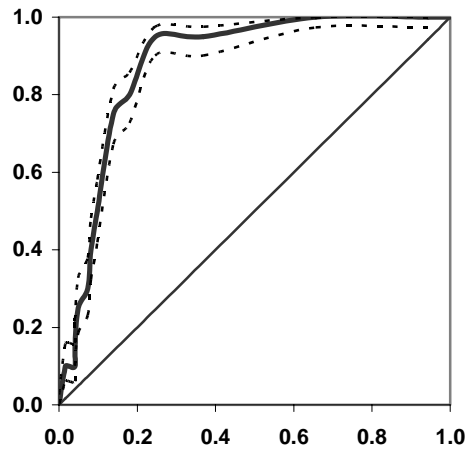
ROC can be represented by a graph of the hit rate against the false alarm rate as  $w$  varies, with the false alarm rate plotted as the  $X$ -axis and the hit rate as the  $Y$ -axis. The location of the entire curve in the unit square is determined by the intrinsic discrimination capacity of the forecasts, and the location of specific points on a curve is determined by the decision threshold  $w$  that is selected by the user. As the decision threshold  $w$  varies from low to high, or the ROC curve moves from right to left,  $H$  and  $F$  vary together to trace out the ROC curve. A perfect discrimination is represented by an ROC that rises from  $(0,0)$  along the  $Y$ -axis to  $(0,1)$ , then straight right to  $(1,1)$ . The diagonal  $H = F$  represents zero skill, indicating that the forecasts are completely non-discriminatory. ROC points below the diagonal represent the same level of skill as they would if they were located above the diagonal, but are just mislabeled, *i.e.*, a forecast of non-occurrence should be taken as occurrence.

Figure 4a: ROC for  $Q0 \pm 95\%$  Band



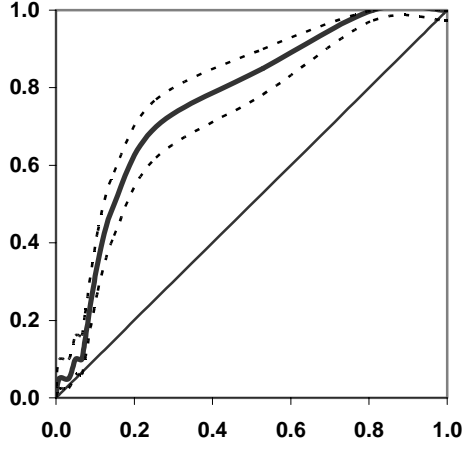
Y-axis: Hit Rate; X-axis: False Alarm Rate

Figure 4b: ROC for  $Q1 \pm 95\%$  Band



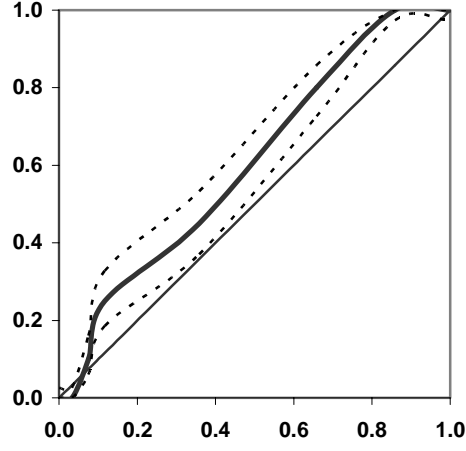
Y-axis: Hit Rate; X-axis: False Alarm Rate

Figure 4c: ROC for Q2 ± 95% Band



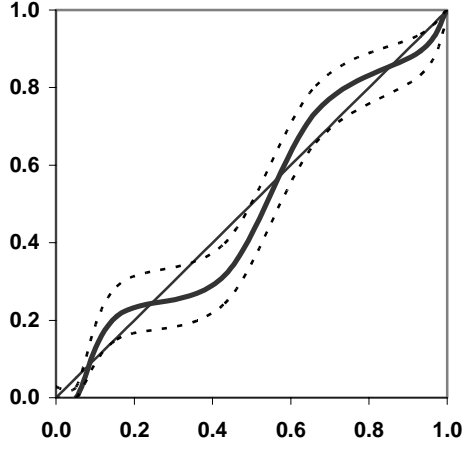
Y-axis: Hit Rate; X-axis: False Alarm Rate

Figure 4d: ROC for Q3 ± 95% Band



Y-axis: Hit Rate; X-axis: False Alarm Rate

Figure 4e: ROC for Q4 ± 95% Band



In Figures 4a-4e the ROC curves together with their 95% confidence intervals for the current quarter and the next four quarters are displayed. The confidence interval was calculated using the formula

$$\frac{\hat{H} + \frac{z_{\alpha/2}^2}{2T} \pm z_{\alpha/2} \left\{ \left[ \hat{H}(1 - \hat{H}) + \frac{z_{\alpha/2}^2}{4T} \right] / T \right\}^{1/2}}{1 + z_{\alpha/2}^2 / T} \text{ for each } w, \text{ where } z_{\alpha/2} = z_{0.025} = 1.96 \text{ for}$$

a standard normal variate.

In situations where the analyst may have only vague idea about the relative costs of type I and type II errors (*e.g.*, in the problem of predicting the turning point in a business cycle), he or she can pick a comfortable hit rate (or false alarm rate) of choice, and the underlying ROC curve will give the corresponding false alarm rate (or hit rate). This will also give an optimal threshold for making decisions.

The hit rates and false alarm rates for selected threshold values in the range 0.50-0.05 are reported in Table 7, where one can find the mix of hit and false alarm rates that are expected to be associated with each horizon-specific forecast.

Table 7: Measures of Forecast Skill: Quarter 0 to Quarter 4

<i>w</i>	<i>Q0</i>				<i>Q1</i>				<i>Q2</i>				<i>Q3</i>				<i>Q4</i>			
	<i>H</i>	<i>F</i>	Kuipers score	Odds Ratio	<i>H</i>	<i>F</i>	Kuipers score	Odds Ratio	<i>H</i>	<i>F</i>	Kuipers score	Odds Ratio	<i>H</i>	<i>F</i>	Kuipers score	Odds Ratio	<i>H</i>	<i>F</i>	Kuipers score	Odds Ratio
0.50	0.55	0.07	0.48	17.57	0.25	0.05	0.20	6.44	0.05	0.03	0.02	1.54	0.00	0.00	0.00	-	0.00	0.00	0.00	-
0.45	0.60	0.07	0.53	19.00	0.30	0.07	0.23	5.38	0.10	0.05	0.05	2.13	0.00	0.00	0.00	-	0.00	0.00	0.00	-
0.40	0.60	0.08	0.52	16.95	0.40	0.08	0.32	7.47	0.10	0.07	0.03	1.57	0.00	0.00	0.00	-	0.00	0.00	0.00	-
0.35	0.70	0.10	0.60	21.58	0.50	0.10	0.40	9.17	0.15	0.07	0.08	2.20	0.00	0.03	-0.03	0.00	0.00	0.00	0.00	-
0.30	0.85	0.11	0.74	44.12	0.75	0.14	0.61	18.53	0.35	0.11	0.24	4.47	0.10	0.08	0.03	1.37	0.00	0.05	-0.05	0.00
0.25	0.90	0.16	0.74	46.35	0.80	0.18	0.62	18.18	0.50	0.15	0.35	5.72	0.25	0.12	0.13	2.52	0.21	0.16	0.06	1.45
0.20	0.95	0.20	0.75	74.48	0.95	0.25	0.70	58.27	0.70	0.26	0.44	6.77	0.45	0.36	0.09	1.47	0.32	0.43	-0.12	0.61
0.15	1.00	0.23	0.77	-	0.95	0.37	0.58	32.51	0.85	0.53	0.32	5.05	0.80	0.66	0.14	2.08	0.74	0.66	0.07	1.42
0.10	1.00	0.40	0.60	-	1.00	0.66	0.34	-	1.00	0.81	0.19	-	1.00	0.86	0.14	-	0.89	0.93	-0.04	0.63
0.05	1.00	0.68	0.32	-	1.00	0.94	0.06	-	1.00	0.99	0.01	-	1.00	1.00	0.00	-	1.00	1.00	0.00	-

Note: *w* = decision threshold; *H* = hit rate; *F* = false alarm rate.

We find, for achieving a hit rate of 90% with Q0 forecasts, one should use 0.25 as the threshold, and the corresponding false alarm rate is expected to be 0.16. Table 7 also shows that at this threshold value, even though the false alarm rates are roughly around 0.15 for forecast of all horizons, the hit rate steadily declines from 90% for Q0 to only 21% for Q4 - clearly documenting the rapid speed of deterioration in forecast capability as the forecast horizon increases. Though not reported in Table 7, for the same hit rate of 90%, the false alarm rates for Q1 through Q4 forecasts are 0.189 ( $w=0.237$ ), 0.636 ( $w=0.13$ ), 0.808 ( $w=0.115$ ) and 0.914 ( $w=0.10$ ) respectively. Thus, for the same hit rate, the corresponding false alarm rates for Q3-Q4 forecasts are so large (80% and 91% respectively) that they can

be considered useless for all practical purposes, and thus, may have very little value in decision-making.

In Table 7 we have also reported the Kuipers scores (KS) and the odds ratios ( $\theta$ ) for selected  $w$ . The rapid decline in these values as the forecast horizon increases is remarkable, and for Q4 forecasts these values are close to zero and unity respectively, suggesting no-skill. Using the critical value 1.645 for a one-sided normal test at the 5% level, the KS and  $\theta$  values were found to be statistically significant for Q0-Q2 and insignificant for Q4 forecasts. For Q3 forecasts, there is some conflicting evidence depending on the tests we use. Based on the standard error formula  $\sqrt{[H(1-H)/(a+c)]+[F(1-F)/(b+d)]}$  for KS reported in Argenti (1996), KS values for Q3 were insignificant at the 5% level for all allowable values of  $w$ . However, the PT test and the test based on log odds ratio for Q3 were statistically significant only for  $w = 0.25$  even at the 1% level. Notwithstanding this result, the weight of our previous evidence suggests that Q3 forecasts have very little skill. We should, however, emphasize that statistical significance or insignificance does not mean the forecasts have utility or value in a particular decision theoretic context.

We find overwhelming evidence that Q0-Q2 forecasts have good operating characteristics. Given the relative costs of two types of classification errors, the end-user can choose an appropriate threshold  $w$  to minimize the total expected cost of misclassification. This type of optimal decision rule cannot be obtained by the Murphy-Yates decompositions of  $QPS$ . More importantly, for forecasting relatively rare business events like recessions, ROC and odds ratios are useful for making sure that the probability forecasts have operational value. This is because, in this approach, the success rate in predicting the predominant event is not part of the goodness of fit measure.

## **CONCLUSIONS:**

We found overwhelming evidence that the shorter run forecasts (Q0-Q2) possess significant skill, and are well calibrated. The resolution or the discrimination ability is also reasonable. Considering the fact that the chronologies of the NBER recessions are usually determined long after the recession is over, negative GDP growth projections are probably a reasonable way of tracking business cycles in real time.

However, the variance of these forecasts, particularly during cyclical downturns, is significantly more than necessary. The analysis of probability forecasts, thus, shows that forecasters respond *also* to cues that are not related to the occurrence of negative GDP growths.

In contrast, Q3 and Q4 forecasts exhibit poor performance as measured by negative skill scores, low resolutions, dismal ROC measures, and insignificant correlations with actual outcomes. Interestingly, the Seillier-Moiseiwitsch and Dawid (1993) test for perfect forecast validity failed to detect any problem with the longer-term forecasts. However, it is clear from our analysis that our professional forecasters do not have adequate information to forecast meaningfully at horizons beyond two quarters; they lack relevant discriminatory cues. Since the SPF panel is composed of professional economists and business analysts who forecast on the basis of models and informed heuristics, their failure for the long-term forecasts may indicate that at the present time forecasting real GDP growth beyond two quarters may not be possible with reasonable type I and Type II errors.

As Granger (1996) has pointed out, in some disciplines forecasting beyond certain horizons is known to be not possible; for instance, in weather

forecasting the boundary seems to be four or five days. Our analysis of probability forecasts suggests that in macro GDP forecasts, two quarters appears to be the limit at the present time.

We have also emphasized that for forecasting rare events, it is important to examine the ROC curves where the relative odds for the event can be studied at depth. The analysis also helps find an optimum probability threshold for transforming the probability forecasts to a binary decision rule. I

Other interesting implications of this study are as follows: First, decomposition methodologies introduced in this paper have much broader implications for evaluating model fit in Logit, Probit and other limited dependent variable models.

