Probability Distributions or Point Predictions? Survey Forecasts of US Output Growth and Inflation

Michael P. Clements

No 976

WARWICK ECONOMIC RESEARCH PAPERS

DEPARTMENT OF ECONOMICS

THE UNIVERSITY OF WARWICK

Probability Distributions or Point Predictions? Survey Forecasts of US Output Growth and Inflation

Michael P. Clements
Department of Economics
University of Warwick*

January 20, 2012

Abstract

We consider whether survey respondents' probability distributions, reported as histograms, provide reliable and coherent point predictions, when viewed through the lens of a Bayesian learning model, and whether they are well calibrated more generally. We argue that a role remains for eliciting directly-reported point predictions in surveys of professional forecasters.

Keywords: probability distribution forecasts, point forecasts, Bayesian learning.

JEL classification: C53.

*I am grateful to Ken Wallis for helpful comments.

1 Introduction

Survey respondents' subjective probability distributions of inflation and output growth are not always consistent with the corresponding point predictions, and when they differ it tends to be in the direction of the point forecasts presenting a more favourable outlook. For the US Survey of Professional Forecasters (SPF), Engelberg, Manski and Williams (2009) and Clements (2009, 2010) examine the relationship between the subjective probability distributions of the individual respondents, which are reported as histograms, and the respondents' point forecasts, for both real output growth and inflation. They conclude that for the majority of cases the two match, but that when they are inconsistent the point forecasts of output growth and inflation tend to suggests a rosier outlook: the output growth and inflation point forecasts are higher and lower, respectively, than measures of central tendency derived from the subjective probability distributions reported as histogram forecasts.[1]

Engelberg *et al.* (2009, p. 40) draw the conclusion that:

'...point predictions may have a systematic, favorable bias. ...agencies who commission forecasts should not ask for point predictions. Instead, they should elicit probabilistic forecasts....'

However, Clements (2010) suggests that the point predictions are more accurate than measures of central tendency derived from the probability distributions, when judged by conventional squared-error loss, casting doubt on the recommendation that surveys need only elicit information on respondents' probability distributions even when 'most likely' outcomes are of interest.

The first conjecture we address in this paper is that professional forecasters taken as a whole are less successful in communicating their best point prediction when they are required to produce a probability distribution or histogram forecast. We suppose they do not use fully-integrated forecasting systems that produce forecast distributions and point forecasts in a mutually consistent fashion, otherwise we ought not observe the inconsistencies documented by Engelberg *et al.* (2009) and Clements (2009, 2010).

Evidence from the direct questioning of survey participants about the forecasting methods or tools that they use suggests that participants 'use a variety of procedures to predict the major expenditure components of GNP, combine these predictions in nominal and real terms, and check and adjust the resulting forecasts for consistency with logic, theory, and the currently available

[1] García and Manzanares (2007) find that the growth and inflation forecasts of the European Central Bank's Survey of Professional Forecasters follow a similar pattern, and Boero, Smith and Wallis (2008b, 2008a) find that the same is generally true of the Bank of England Survey of External Forecasters, especially for the output growth forecasts.

information' (as summarised by Zarnowitz and Braun (1993, p. 23)). Zarnowitz and Braun go on to state that forecasters rarely rely on a single forecasting method or model, and usually draw on a range of sources to inform their forecasts (including econometric models, leading indicators, and anticipations surveys) as well as exercising their own judgment. Batchelor and Dua (1991) found that 'judgment' was cited as being the single most important forecasting technique by 51% of the Blue Chip Panel, with 28% reporting econometric modelling, and 21% time series analysis.

Given the use of different methods and the vagaries of the application of judgment at different levels, it is perhaps not surprising that the different ways of expressing forecasts of the same object do not always match. And that forecasters may be better able (or have more incentive) to communicate their 'best' or 'most likely' forecasts as single point predictions, rather than as implied central tendencies of histograms.[2]

Our second conjecture is that the fault may lie with the econometrician. When probability distributions are elicited, it is typically in the form of a histogram. Generally this provides insufficient information for the purpose at hand, such as calculating moments and probability integral transforms. We consider whether the ways of estimating probability distributions from histograms used in the literature may fail to do justice to the underlying subjective distributions. We propose a way of doing so based on an assessment of those histograms which provide near error-free estimates for a probability-integral-transform approach.

The evidence we provide in support of the first conjecture is based on fitting the Bayesian learning model (BLM) jointly to each individual respondent's histogram mean forecasts (derived from the respondents' histograms using the approach favoured by Engelberg *et al.* (2009)) and their point predictions. We then develop a formal test of whether the respondents update both types of forecast in the same way as the forecast horizon shortens. This is our first methodological contribution. The BLM offers a simple description of how forecasts should be updated as new information becomes available. If the estimates of the histogram means constitute a coherent set of forecasts, we would expect that more weight would be given to new information when the forecast horizon is short, relative to when it is long. At long horizons the respondents' beliefs about the

[2]The literature on the psychology of judgement under uncertainty recognises that there may be differences. For example, O'Hagan, Buck, Daneshkhah, Eiser, Garthwaite, Jenkinson, Oakley and Rakow (2006, ch. 3) provides a concise review of Kahneman and Tversky's 'heuristics and biases' research programme, which suggest that probability assessments are made using limited information and 'quick-and-easy' shortcuts. Further, their review of Hammond's cognitive continuum theory (O'Hagan *et al.* (2006, p.56)) suggest that task characteristics may matter, and specifically that intuitive as opposed to analytical thinking may be encouraged when 'minimal feedback is obtained, and high accuracy is not expected'. The individuals' histograms are less amenable to accuracy assessment, and comparison one to another, than the point forecasts.

expected long-run mean growth rate should hold sway, with current developments becoming more influential as the horizon shortens. The estimates of the BLM key parameters allow a simple assessment of whether the forecasts conform to these fairly minimal requirements. Manzan (2011) has recently estimated such a model for the point predictions, and found that these properties hold, albeit that there is heterogeneity in the model estimates across individuals. When we jointly model an individual's histogram means and point predictions we find that the evolution of the means of the histograms are not well explained by the BLM, contrary to the findings for the point predictions. Secondly, we assess whether the histograms are accurate in the sense of being 'correctly calibrated', and so go beyond an assessment of the first-moment properties.

With regard to the second conjecture, we are aware that estimates of the means of the histograms will depend on the distributional assumptions we make, as will our assessments of whether they are correctly calibrated. Hence we undertake sensitivity exercises - we present results for a number of distributional assumptions to assess whether the findings are sensitive to the assumptions we make. By and large, our qualitative findings are not sensitive to the distributional assumptions. It turns out that our proposed method of determining whether the underlying subjective distributions are 'well calibrated', in the sense that the forecast probabilities are close to the actual probabilities, is not definitive because of certain characteristics of the sample of forecasts and outturns, as explained below. However, our novel way of testing the the underlying subjective distributions directly (that is, free of any additional assumptions about the distribution of probability mass within a histogram interval) might prove useful in other applications.

Although doubts remain as to how well we approximate the underlying subjective distributions, our general conclusion is that the use of best-practice methods to calculate continuous distributions from histograms results in poorly calibrated probability distributions and gives estimates of means that turn out to be relatively inaccurate compared to point predictions. This suggests there is a case for eliciting point predictions irrespective of whether the underlying subjective distributions, or the distributional assumptions made by the econometrician, are at fault.

The remainder of the paper is organised as follows. Section 2 describes the SPF forecast data, the calculation of means and the fitting of continuous distributions to the histograms. Section 3 outlines the Bayesian learning model, which will be used to contrast the means of the individuals' probability distributions and their point predictions. Section 4 provides the analysis of whether the individual probability distributions are well calibrated, using the probability integral transform approach, and section 5 considers whether these findings are affected by the reported histograms only partially

revealing the underlying subjective probability distributions. Section 6 offers some concluding remarks. Finally, an appendix outlines the calculation of point predictions and histogram means for a single respondent's returns to two adjacent surveys. These forecasts were not selected randomly, but were chosen as an example of the way in which the two types of forecasts are clearly not generated as part of an integrated forecasting system. These forecasts are not meant in any sense to be typical: our case against the use of histogram forecasts for point prediction rests on the empirical findings reported in sections 3 and 4, not on this anecdotal evidence. These detailed calculations are meant to better illuminate the nature of the calculations that underpin the results reported in these sections. However it does serve to show that not all professional forecasters are as careful in their deliberations as is sometimes assumed.[3]

2 Description of data and the calculation of means and parametric distributions for the SPF histogram forecasts

We choose the SPF as our source of survey expectations because it contains information on respondents' probability distributions for inflation and output growth as well as their point forecasts for these key macro-aggregates, and spans a reasonably long historical period. It is a quarterly survey of professional forecasters of the US economy. The SPF began as the NBER-ASA survey in 1968:4 and runs to the present day: see e.g., Croushore (1993). The last quarter we use is for the fourth quarter of 2010, giving 169 quarterly surveys of expectations data spanning the last 40 years.[4]

The histograms are of the percentage change in the survey year relative to the previous year. We calculate matching year-on-year point forecasts as follows. The surveys provide point forecasts of the level of the variable in the current (survey quarter) and each of the next four quarters. We use the forecasts of the current and subsequent quarters, along with the actual values from the vintage of data available at that time to calculate the forecast of annual inflation or output growth

[3] It is sometimes argued that surveys of *professional* forecasters, as distinct from surveys of 'lay people', are more likely to provide meaningful, informed responses, as professional forecasters are more knowledgeable and likely to respond to incentives (e.g., reputations).to report accurately.

[4] This was downloaded on 15th September 2011, and thus includes the corrections released on August 12, 2011. Further details are available at http://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/

for the current year over the previous year.[5,6] So for Q1 surveys, we sum the forecast of the current quarter and the forecasts of the next three quarters,[7] and divide by the data for the previous year's four quarters. For Q2 surveys the approach is the same except the value for the preceding quarter (Q1) is now data, and similarly for surveys made in the third and fourth quarters of the year. So we have forecasts of annual inflation made in Q1 through to Q4 of that year.

We analyse the mean forecasts of output growth and inflation implied by the histograms, and for section 4 we will also require continuous distribution approximations to the histograms. To make matters concrete, the first two columns of table 1 illustrate a typical SPF histogram forecast, where the respondent has assigned probabilities to the given intervals. The open-ended intervals contain the probabilities attached to inflation (say) being less than -2, and greater than 6, respectively (in our example these happen to both be zero). These intervals are closed by replacing '< -2' and '6+' by '-3 to -2.1' and '6 to 6.9'. Following Engelberg *et al.* (2009), we assume that the histogram provides the cdf points 'x' (recorded in the third column of the table) with the associated probabilities in the fourth column of the table. We maintain these assumptions throughout - when we fit parametric distributions to the histograms, and when we calculate means directly from the histograms.

For the direct calculation of means from the histograms, we obtain the same result whether we assume that the probability mass is uniform within an interval or all at the interval midpoint. Alternatively, we can first fit a distribution to the histograms: Giordani and Söderlind (2003, p. 1044) fit normal distributions to the histograms[8], and the generalized beta distribution is also a popular choice (see, e.g., O'Hagan *et al.* (2006), and the application to the SPF histograms by Engelberg *et al.* (2009)). If the distribution underlying the histogram is approximately 'bell-shaped'

[5] We use the quarterly Real Time Data Set for Macroeconomists (RTDSM) maintained by the Federal Reserve Bank of Philadelphia: see Croushore and Stark (2001). This consists of a data set for each quarter that contains only those data that would have been available at a given reference date: subsequent revisions, base-year and other definitional changes that occurred after the reference date are omitted. Hence we can re-create the annual growth rate forecasts that the respondents would have made had they been asked to report these.

[6] Both the definition and base year of output and the price deflator have changed over time. The vintages of data in the RTDSM match the indices for which probability assessments and point forecasts were requested in the SPF, so that these changes are inconsequential for our use of the survey data. The only problem is that prior to 1981:3, the output growth histogram question related to nominal GNP. Hence for output growth the sample of forecasts we study is restricted to the surveys from 1981:3 onwards.

[7] As of 1981:3, forecasts of the levels of the variables for the current year were recorded. However, summing the quarterly forecasts allows us to use data back to 1968:4 for inflation.

[8] Fitting normal approximations requires two or more intervals with non-zero probabilities attached. This provides a minimum of three points on the gaussian cdf, which uniquely identify the mean and variance. For single interval histograms (i.e., a probability of 1 assigned to one of the intervals) we take the mean to be the mid-point of that interval.

then the uniformity assumption will tend to overstate the dispersion of the distribution. Moreover if there is a large difference in the probability mass attached to adjacent intervals, then it might be desirable to attach higher probabilities to points near the boundary with the high probability interval, which will be facilitated by fitting a parametric distribution. Engelberg *et al.* (2009) argue in favour of the unimodal generalized beta distribution over the normal distribution as a way of allowing for asymmetry in the individual's assessments of risks. We follow their approach, and fit unimodal generalized beta distributions when non-zero probabilities are assigned to three or more histogram intervals, and also follow their recommendations for histograms with a single interval or two non-zero intervals (see p.37-8 for details of the fitting methods). The estimates of the means are then calculated from the parameters of the fitted distributions.

3 Model of Bayesian learning and differential interpretation of information

Our learning model is based on Kandel and Zilberfarb (1999), as adapted by Manzan (2011). We let $F_{it,h+1}$ denote the forecast by individual i at period $t-h-1$ of Y_t. In terms of the standard BLM, this forecast is assumed to correspond to the mean of a gaussian prior distribution, with variance $a_{it,h+1}^{-1}$. At time $t-h$, all individuals receive a common public signal, L_{th}, about Y_t, and based on the signal and their prior they report a new forecast, their posterior forecast. The Kandel and Zilberfarb (1999) model allows that individuals may interpret the signal differently. Allowing that individuals 'have differing beliefs about the distribution of signals' - the *differential-interpretation hypothesis* means that some believe the signal mean is above the mean of the distribution of Y_t, while others will interpret the signal as an under-estimate. This is modelled by assuming that individual i's best guess of Y_t, based on the signal alone, is given by $Y_{ith} = L_{th} - \varepsilon_{ith}$, where $\varepsilon_{ith} \sim N\left(\mu_{ith}, b_{ith}^{-1}\right)$. When $\mu_{ith} > 0$, the signal is interpreted as likely to over-estimate Y_t, and when $\mu_{ith} < 0$ the signal is expected to under-estimate the outcome.

By Bayes rule and the assumed normality of the prior and the likelihood, the optimal posterior forecast is:

$$F_{ith} = \lambda_{ith} F_{it,h+1} + (1 - \lambda_{ith}) (L_{th} - \mu_{ith}), \tag{1}$$

where $\lambda_{ith} = a_{it,h+1} (a_{it,h+1} + b_{ith})^{-1}$, the ratio of the precision of the prior to the precision of the

posterior forecast. These are the optimal weights to attach to the prior and the signal. Given two forecasts $F_{it,h+1}$ and F_{ith} of the same target Y_t, we can interpret $F_{it,h+1}$ as the prior and F_{ith} as the posterior. Equation (1) is typically augmented with a reporting error v_{ith}, where $v_{ith} \sim N\left(0, \sigma_v^2\right)$, so that:

$$F_{ith} = \lambda_{ith} F_{it,h+1} + (1 - \lambda_{ith}) (L_{th} - \mu_{ith}) + v_{ith}. \tag{2}$$

Kandel and Zilberfarb (1999) construct a test of the differential-interpretation hypothesis based on the cross-sectional variances of the prior and posterior forecasts. A number of recent papers follow Kandel and Zilberfarb (1999) in considering the implications of the hypothesis for the cross-sectional variance of survey point forecasts, or disagreement.[9]

At this point, we follow Manzan (2011). Rather than considering the cross-sectional dispersion, we fit the BLM given by (2) directly to the individual-level forecast data. This allows us to estimate the behavioural parameters of (2) on the observed (forecast) data once we have a measure of L_{th}. The key parameters are allowed to differ by forecast horizon h, and individual i. One would expect the weights given to the signal and prior to vary with h, for the reasons given by Lahiri and Sheng (2008, 2010) and Patton and Timmermann (2010), *inter alia*. At long horizons, prior beliefs about the long-run means of the variables under study are likely to dominate. Assuming stationarity (of the output growth rate, and the rate of inflation), the current state of the economy will affect the short-term outlook but will be far less informative about longer-term developments. Hence greater weight would be expected to be placed on the signal as the horizon shortens. Manzan (2011) tests homogeneity across individuals in terms of (2) as the null that: $\lambda_{ih} = \lambda_h$ and $\mu_{ih} = \mu_h$. Manzan fits the model to the SPF point predictions, and finds evidence against the null. Consequently, we allow as part of our maintained model that the weights and interpretation parameters are both horizon and respondent specific.

Our aim is to assess how well the BLM explains the histogram mean data. There are two aspects to this. Firstly, we informally compare the BLM parameter estimates for the histogram mean data with the estimates for the point prediction. We do this for those respondents who made a sufficient number of forecasts of both types. Secondly, we test whether the BLM parameters are the same

[9] Lahiri and Sheng (2008, 2010) use a Bayesian learning model to investigate the relative importance of the different factors contributing to disagreement as the forecast horizon changes, in a fixed-event forecasting environment. Patton and Timmermann (2010) also estimate a model of cross-sectional dispersion which allows agents to receive different signals about the unknown state of the economy, and to have different priors about the long-run mean values of the two variables they consider, inflation and output growth. Heterogeneity in forecasters' information sets is found to be relatively unimportant in explaining cross-section dispersion, while heterogeneity in priors plays an important role.

across the two types of forecast for each respondent. In both cases we fit the model simultaneously to the point forecast and histogram mean forecasts of each of the survey respondents. We generalise our notation such that a superscript $^{(1)}$ denotes a point forecast, or a parameter related to the point forecast model, and a $^{(2)}$ a histogram mean. Hence (2) becomes the system of two equations:

$$\begin{aligned} R_{ith}^{(1)} &= \delta_{ih}^{(1)} + \alpha_{ih}^{(1)}\left(L_{th} - F_{it,h+1}^{(1)}\right) + v_{ith}^{(1)} \\ R_{ith}^{(2)} &= \delta_{ih}^{(2)} + \alpha_{ih}^{(2)}\left(L_{th} - F_{it,h+1}^{(2)}\right) + v_{ith}^{(2)} \end{aligned}$$

where $R_{ith}^{(j)} = F_{ith}^{(j)} - F_{it,h+1}^{(j)}$ denotes the revision in the point forecast ($j = 1$) or in the histogram mean ($j = 2$), $\alpha_{ith}^{(j)} = \left(1 - \lambda_{ith}^{(j)}\right)$ is the weight on the signal (less the prior forecast $F_{it,h+1}^{(j)}$ and adjusted for the interpretation parameter $\mu_{ith}^{(j)}$). The intercepts are $\delta_{ih}^{(j)} = -\mu_{ih}^{j}\alpha_{ih}^{(j)}$ for $j = 1, 2$. The error term incorporates both reporting errors and any non-observed part of the signal not captured by L_{th}. We suppose the signal for the histogram mean and for the point forecasts is the same. Hence L_{th} is not super-scripted. We have dropped the time t subscripts from the parameters in order to indicate that these are assumed not to vary over time.

We can write the model more succinctly for all individual respondents $i = 1, 2, \ldots, T_i$ as:

$$R_h = X_h\beta_h + v_h \tag{3}$$

where $R_h = \left[R_{1h}', \ldots R_{T_i,h}'\right]'$, $R_{ih} = \left(R_{ih}^{(1)\prime}\ R_{ih}^{(2)\prime}\right)'$, $R_{ih}^{(j)} = \left(\ldots R_{ith}^{(j)} \ldots\right)'$, $j = 1, 2$. [10] Hence R_h contains all the revisions to point forecasts of the first respondent, followed by all the revisions to their histogram means, and then repeats for the second, and then the third, respondent, and so on, up to the T_i'th respondent. The explanatory variables are similarly ordered:

$$X_h = \begin{bmatrix} x_{1h}^1 & & & & \\ & x_{1h}^2 & & & \\ & & \ddots & & \\ & & & x_{T_i,h}^1 & \\ & & & & x_{T_i,h}^2 \end{bmatrix}$$

[10] We only use observations t which are common to both $R_{ih}^{(1)}$ and $R_{ih}^{(2)}$, but generally respondents who provide point forecasts at t also provide histograms too, and vice versa.

where x_{ih}^1 and x_{ih}^2 contain the T_{n_i} rows of observations on the RHS variables for individual i's point and histogram mean forecasts, respectively. These RHS variables consist of a vector of 1's, then either $\left(\ldots L_{th} - F_{it,h+1}^{(1)} \ldots\right)'$ or $\left(\ldots L_{th} - F_{it,h+1}^{(2)} \ldots\right)'$, and are augmented in each case by the cross-sectional averages of the dependent variable and the 'slope' explanatory variable. The cross-sectional averages are included, as in Manzan (2011), to account for the effects of unobserved factors that impinge on all respondents in a similar fashion at a given time: this is the common correlated effects estimator of Pesaran (2006). However, because we have a system of two equations for each individual, it seems reasonable to also allow for correlated idiosyncratic errors for the two forecasts made at the same point in time. In terms of the $2T_n$ vector of disturbances v_t, organised consistently with the data vectors as:

$$v_h = \left(v_{11h}^{(1)}, v_{12h}^{(1)}, \ldots v_{1,T_{n_1},h}^{(1)}; v_{11h}^{(2)}, v_{12h}^{(2)}, \ldots v_{1,T_{n_1},h}^{(2)}; \ldots; v_{T_i1h}^{(1)}, v_{T_i2h}^{(1)}, \ldots v_{T_i,T_{n_{T_i}},h}^{(1)}; v_{T_i1h}^{(2)}, v_{T_i2h}^{(2)}, \ldots v_{T_i,T_{n_{T_i}},h}^{(2)}\right)'$$

we impose the following structure. We suppose that $E\left(v_{ith}^{(j_1)} v_{ksh}^{(j_2)}\right) = 0$ when either $t \neq s$ or $i \neq k$. When $t = s$ and $i = k$, $E\left(v_{ith}^{(j_1)} v_{ith}^{(j_2)}\right) = \sigma_{ih}^{2\,(1)}$ when $j_1 = j_2 = 1$; $= \sigma_{ih}^{2\,(2)}$ when $j_1 = j_2 = 2$, and $= \sigma_{ih}^{(12)}$ when $j_1 \neq j_2$. We estimate (3) by OLS (for each h), and then use the two-step GLS estimator:

$$\hat{\beta}_h = \left(X_h' \hat{\Omega}^{-1} X_h\right)^{-1} X_h' \hat{\Omega}^{-1} R_h$$

where $\Omega = E\left(v_h v_h'\right)$ is replaced by an estimator constructed from the OLS residuals in accordance with the covariance structure defined above.

We have used T_i to denote the number of forecasters, T_{n_i} the number of forecasts by i, so that $T_n = \sum_{i=1}^{T_i} T_{n_i}$ is the total number of forecasts of either type. Consequently, the number of rows of R_h and X_h is $2T_n$, and the column dimension of X_h is $8T_i$ (we are freely estimating 8 parameters for each of the T_i individuals - 4 for each of their point and histogram mean forecasts).

Our focus is on whether for each individual the weight and interpretation parameters are equal across the two types of forecasts. This contrasts to Manzan (2011) who tests whether the weight and interpretation parameters are identical across individuals for the point forecasts. The hypothesis that each individual forecaster gives the same weight to news when they update their point forecast as they do when they update their histogram mean forecast, when the forecast horizon is h, is given by:

i) Equal weights on news. $H_0 : \alpha_i^{(1)} - \alpha_i^{(2)} = 0$, $i = 1, 2, \ldots, T_i$.[11]

The hypothesis that each forecaster interprets the news in the same way when they update their point and histogram mean forecasts is:

ii) Equal interpretation of news. $H_0 : \mu_i^{(1)} - \mu_i^{(2)} = 0$, $i = 1, 2, \ldots, T_i$. In terms of the estimable parameters, $\mu_i^{(1)} - \mu_i^{(2)} = 0 \Rightarrow \frac{\delta_i^{(1)}}{\alpha_i^{(1)}} - \frac{\delta_i^{(2)}}{\alpha_i^{(2)}} = 0$.

The two hypotheses of interest are tested using Wald-type tests. That is, the test statistic takes the form:

$$f\left(\hat{\beta}_h\right)' \left[R\left(\hat{\beta}_h\right)\left(X_h'\hat{\Omega}X_h\right)^{-1} R\left(\hat{\beta}_h\right)'\right]^{-1} f\left(\hat{\beta}_h\right) \tag{4}$$

which is distributed $\chi^2_{T_i}$ under the null. The test for equal weights is a linear hypothesis, and $f(\beta_h)$ specialises to $f(\beta_h) = R\beta_h$, where R is T_i by $8T_i$, with typical i^{th} row given by

$$R_{i\cdot} = \left(0_{1\times(i-1)8},\ 0\ 1\ 0\ 0\ 0\ -1\ 0\ 0,\ 0_{1\times(T_i-i)8}\right). \tag{5}$$

The parameter vector β_h is ordered as:

$$\beta_h = \left(\delta_1^{(1)}, \alpha_1^{(1)}, *, *, \delta_1^{(2)}, \alpha_1^{(2)}, *, *; \ldots; \delta_{T_i}^{(1)}, \alpha_{T_i}^{(1)}, *, *, \delta_{T_i}^{(2)}, \alpha_{T_i}^{(2)}, *, *\right)'$$

where the '$*$'s denote coefficients on the cross-sectional averages of the dependent and the explanatory variables. Hence for the i^{th} forecaster, the form of (5) gives $R_{i\cdot}\beta_h = 0 \Rightarrow \alpha_i^{(1)} - \alpha_i^{(2)} = 0$, i.e., equality of the weights on the signal in the equations for an individual's point and histogram mean forecasts.

For the hypothesis of equal interpretation effects,

$$f(\beta_h) = \begin{bmatrix} \frac{\delta_1^{(1)}}{\alpha_1^{(1)}} - \frac{\delta_1^{(2)}}{\alpha_1^{(2)}} \\ \frac{\delta_2^{(1)}}{\alpha_2^{(1)}} - \frac{\delta_2^{(2)}}{\alpha_2^{(2)}} \\ \vdots \\ \frac{\delta_{T_i}^{(1)}}{\alpha_{T_i}^{(1)}} - \frac{\delta_{T_i}^{(2)}}{\alpha_{T_i}^{(2)}} \end{bmatrix}$$

[11] We have dropped the h-subscript from the elements of β_h for simplicity.

and $R(\beta_h) = \frac{\partial f(\beta_h)}{\partial \beta_h'}$ has typical row:

$$\frac{\partial \left[f(\beta_h)_{i.} \right]}{\partial \beta_h'} = \left[0_{1\times(i-1)8}, \left(\frac{1}{\alpha_i^{(1)}} \quad \frac{-\delta_1^{(1)}}{\alpha_i^{(1)\ 2}} \quad 0 \quad 0 \quad \frac{1}{\alpha_i^{(2)}} \quad \frac{-\delta_2^{(2)}}{\alpha_i^{(2)\ 2}} \quad 0 \quad 0 \right), 0_{1\times(T_i-i)8} \right].$$

where $f(\beta_h)_{i.}$ denotes the i^{th} row of $f(\beta_h)$.

3.1 Fitting the Bayesian learning model to the SPF data

We estimate (3) for three values of h, $h = 0, 1, 2$, where $h = 0$ defines a forecast of the annual growth rate made in a fourth-quarter survey. When $h = 0$, we calculate the vector of forecast revisions R_{i0} as the difference between the fourth-quarter and preceding third-quarter survey forecasts; $h = 1$ uses the third and second-quarter survey forecasts; and $h = 2$ the second and first-quarter survey forecasts. We use only those individuals for which we have at least 10 revisions for each type of forecast at a given horizon, i.e., $T_{n_i} \geq 10$.

Estimation of (3) requires that we specify the signal at period $t - h$, L_{th}. While it seems reasonable to assume that this is the same for all individuals, in the sense that private information would be expected to be largely unimportant for macro-aggregates such as output growth and inflation, there is nevertheless a potentially vast amount of economic and financial data that might be expected to be informative about the future course of the economy. The solution proposed by Manzan (2011) is simply to use the latest estimate of Y available at period $t - h$, on the grounds that this should be a good predictor of Y_t. Given the delay of one quarter in the publication of national accounts data, the latest data available at $t - h$ will be the advance estimate of Y in the previous period: we denote this by Y_{t-h-1}^{t-h}, where the superscript denotes the data vintage and the sub the observation time period.[12] That part of the signal not captured by $L_{th} = Y_{t-h-1}^{t-h}$ is assumed to affect the revision to the forecast between $t - h - 1$ and $t - h$ via the error term v_{ith} in (3), and we use the estimator proposed by Pesaran (2006) to allow for correlated effects across respondents. As a check that our results are not being driven by this choice for the signal, we also generate the signal as a model-based forecast of the target given data up to and including Y_{t-h-1}^{t-h}.

[12] Note that the use of the latest-available data observation as the signal corresponds to a 'no-change' predictor. At least for inflation, there is evidence to suggest that such a predictor offers competitive forecasts. Atkeson and Ohanian (2001) show that a random walk model for US inflation is generally as good as models that use 'activity variables' as explanatory variables (as in Phillips Curve forecasting models). See also Stock and Watson (2007, 2010) for recent assessments. However, a variable such as output growth (first-order autocorrelation coefficient of around 0.3) is less persistent, and the no-change forecast may be dominated by a simple autoregression, other things being equal.

We use forecasts from autoregressive forecasting models for the quarterly percentage growth rates, using an estimation period that includes data through $t-h-1$ from the $t-h$ data vintage.[13]

Our coverage of the results begins by plotting the estimates of $\left\{\mu_{ih}^{(1)}, \mu_{ih}^{(2)}, \alpha_{ih}^{(1)}, \alpha_{ih}^{(2)}\right\}$ obtained from GLS estimation of (3) for each respondent for $h = 0, 1, 2$: see figures 1 to 4. The first two figures are for output growth, and figures 3 to 4 are for inflation. Figures 1 and 3 are for L_{th} equal to the model forecast of Y_t, and figures 2 and 4 are for $L_{th} = Y_{t-h-1}^{t-h}$. Within each figure, the left-hand-side panels refer to the weight parameter and the bias parameter for the point forecasts, those on the right are for the histogram mean forecasts. Within each panel, the estimates for a given individual for each of the three forecast horizons are joined up.

For output growth, the estimates of the weight parameter for the point forecasts $\left\{\alpha_{ih}^{(1)}\right\}$ indicate a reduction in dispersion and more weight being placed on the signal as h shortens. Forecasters regard the signal as being more informative at short horizons. Secondly, the interpretation parameter for the point forecasts is for most respondents relatively small, especially at $h = 0$, when the estimates $\mu_{i0}^{(1)}$ are tightly clustered about zero. These findings hold for both definitions of the signal (figures 1 and 2).

The findings for the inflation point forecasts are broadly similar.[14] By and large, we find reasonable estimates of the parameters of the Bayesian learning model when fitted to the individual-level point forecasts for inflation and output growth. This was to be expected given the results in Manzan (2011).

However, we do not obtain readily-interpretable estimates of the BLM when it is estimated on the histogram mean forecasts for either variable. There is little evidence that the bias parameter gets smaller as h shortens for inflation, and remains as large as ± 0.4 percentage points at $h = 0$. For output growth the values of this parameter are more dispersed than for the point predictions at $h = 0$. There is less evidence that the weight on news increases as the horizon shortens, relative to the case for the point predictions, for both variables. In short, neither the histogram mean forecasts of output growth or of inflation appear to constitute coherent sets of forecasts when viewed through

[13] The AR model orders are estimated on rolling windows of 40 observations, selecting the model order at each instance by BIC. To illustrate, when $h = 2$, we have a Q2 survey. This means that we have the level of Y up to Q1. A model is estimated on quarterly growth rates, and 1 to 3 step ahead forecasts are generated iteratively using actual data up to Q1. These forecast quarterly growth rates are used to calculate forecasts of the level of Y in Q2, Q3 and Q4. This enables us to construct the 'signal' – the expected annual year-on-year growth rate using the Q2 vintage data.

[14] For inflation a few 'rogue' parameter value estimates are excluded from the figures to aid their interpretability. For the $h = 2$ forecasts we obtained a few estimates of α close to zero, which translated into very large μ values relative to the range of the majority of the points plotted in the figures. (Recall that the μ's are the ratio of the intercepts to the α parameters).

the lens of the BLM. When forecasters update their histogram mean forecasts they do not appear to do so in a way which is consistent with the BLM.

Next, we turn to the formal tests of the equivalence of the BLM weight and interpretation parameters for each individual. Recall that we are not testing whether the key parameters are the same across all respondents, as in Manzan (2011), but whether they are equal across the two types of forecast for a given individual. The results are shown in table 2. The first panel records the results for our preferred method of estimating means - from fitting generalized beta distributions as in Engelberg *et al.* (2009) to allow asymmetry in the underlying subjective probability distributions.

We find that we can reject the hypothesis that every respondent applies the same weight to news when they revise their point forecast and histogram mean predictions for the shortest horizon revision $h = 0$ for both output growth and inflation, and also for $h = 1$ for output growth. These results hold irrespective of how the signal is defined. We also reject the null of equal interpretation bias for $h = 1$ for output growth forecasts (using the lagged actual as the signal). There is perhaps less formal evidence against the null hypotheses than might be expected given the differences in the estimates in the figures. This is likely to be due to low power given the relatively small samples of forecasts for a given individual at a particular horizon.

The second and third panels are the sensitivity check: that the results are not solely due to fitting the generalised beta distiribution to estimate mean forecasts. The second panel records results for calculating means in the 'standard way', which assumes the probability mass in each histogram interval is located at the midpoint of the interval. In the bottom panel the means are estimated from gaussian distributions fitted to the histograms as in Giordani and Söderlind (2003).

The results are largely unchanged in each case: the null that each individual applies the same weight to news when they revise their point predictions, as when they revise their histogram means, is again rejected for both variables at $h = 0$, and for output growth for $h = 1$.

4 Probability integral transform evaluation of the individual forecast histograms

The recent popularity of density forecasts in the economics and finance forecasting literature stems from the fact that they provide more information than an estimate of the central tendency (see Tay and Wallis (2000), and Hall and Mitchell (2009) for surveys). A now standard way of evaluating density forecasts is the probability integral transform (pit) approach popularized by Diebold, Gun-

ther and Tay (1998), which evaluates the whole density, rather than specific moments derived from the density. The approach requires a continuous distribution function, so the results we obtain may in general depend on the assumptions we make in calculating distributions from the reported histograms. Fortunately, for a subset of the forecast data, the histograms are fully informative for the pit-evaluation of those histograms. In section 5 we consider this subset of forecasts, and what it indicates about *i*) the accuracy of the histograms in general and *ii*) the validity of our parametric approximations to the histograms.

We begin with an analysis of all the histograms. If we let the h-step ahead forecast density for the value of a random variable $\{Y_t\}$ be denoted by $p_{Y,t|t-h}(y)$, then the probability integral transform (pit) is the forecast probability of Y_t not exceeding the realized value y_t:

$$z_t = \int_{-\infty}^{y_t} p_{Y,t|t-h}(u)du \equiv P_{Y,t|t-h}\left(y_t\right). \tag{6}$$

When the forecast density equals the true density, $f_{Y,t|t-h}\left(y\right)$, it is simple to show that $z_t \sim U\left(0,1\right)$. When $h = 1$, and given a sequence $t = 1, \ldots, n$, the time series $\{z_t\}_{t=1}^{n}$ is independently identically uniform distributed, i.e., *iid* $U\left(0,1\right)$, under the null that at each t the forecast and true densities match. We can obtain non-overlapping series of forecasts – in the sense that the realized value is known before the next forecast is made – by treating separately the density forecasts made in a given quarter of each year. This avoids the counterpart of the well-known problem in the point forecast evaluation literature, whereby a sequence of optimal h-step ahead forecasts (with forecasting interval of one period) will follow a moving-average process of $h-1$. Hence we are able to evaluate the series of pits as if they were from one-step ahead forecast densities. We take as actual values the data released in the second quarter of the subsequent year.[15]

The difficulties in calculating moments from histograms discussed in section 2 extend to the calculation of pits from histograms. We calculate pits after fitting generalized beta distributions, normal distributions, and by assuming probability mass is uniform within a interval. The last is best explained by an example. Suppose the actual value is $y = 3.6$. For the example in table 1, the

[15] The first estimate of the annual growth rate available in the first quarter of the subsequent year would include an advance estimate of the last quarter of the year. Our choice is inkeeping with the literature which generally uses BEA 'final' estimates in preference to advance estimates. Note we do not use the latest-available data vintage (at the time of the study), as this will include benchmark revisions and annual revisions (see, e.g., Fixler and Grimm (2005, 2008) and Landefeld, Seskin and Fraumeni (2008)).

probability integral transform is:

$$\Pr\left(Y < y = 3.6\right) = F\left(3\right) + \frac{y-3}{1}(F\left(4\right) - F\left(3\right)) = 0.5 + 0.6 \times 0.2 \tag{7}$$

(for the earlier histograms with interval widths of two percentage points, the denominator in the above expression is 2).

Table 3 contains the evaluation of the individuals' histograms. Rather than assessing whether the individual sequences, say, $\{z_t\}_{t=1}^{T_{n_i}}$ are *iid* $U\left(0,1\right)$, we follow the suggestion of Berkowitz (2001) and assess whether the inverse normal CDF transformation of the $\{z_t\}$ series (say, $\{z_t^*\}$) is *iid* $N(0,1)$.[16] Berkowitz argues that more powerful tools can be applied to testing a null of *iid* $N(0,1)$, compared to one of *iid* uniformity. We calculate a three-degree of freedom likelihood ratio test of zero-mean, unit variance and independence (specifically, zero first-order autocorrelation) using gaussian likelihood functions. The assumption of normality of $\{z_t^*\}$ is also amenable to testing, and we calculate the Doornik and Hansen (1994) tests of normality. For each respondent with a minimum of 10 responses at a given horizon, we calculate these two tests of their transformed pits. The table reports the proportion of individuals for which we reject the null for each of the two tests at the 10% significance levels.

As is evident from the table, there are differences between the three panels, and across survey horizons, but taken together the histogram forecasts are rejected for around a half of the survey respondents on one or other of the tests of the pits. This is surprising given that these tests are expected to have low power given the relatively small numbers of forecasts by an individual of a given horizon. It is generally true that calculating pits by linear interpolation (middle panel) leads to fewer rejections than the two parametric distribution methods. We conclude that there are question marks about the accuracy of individuals' probability assessments. Because the test rejections are not readily informative about the ways in which the histograms are deficient, we also present box plots of the z's (not the z^*'s) for each respondent (for a given h): see figures 5 and 6 for output growth, and figures 7 and 8 for inflation. Under the null of correct specification, the z's are $U\left(0,1\right)$, so the box which denotes the interquartile range should be approximately positioned between 0.25 and 0.75, with the median (depicted as the horizontal line within the box) close to 0.5. Too small (large) a box (which is correctly centred on 0.5) indicates too much

[16] Values of z equal to 0 or 1 are problematic when we calculate $z^* = \Phi^{-1}\left(z\right)$, where Φ is the gaussian cdf. Note that $z = 1$ when linear interpolation is used and the actual is above the highest interval to which a non-zero probability is attached, and similarly for $z_t = 0$. We set values of z of 0 and 1 to 0.01 and 0.99, respectively.

(little) probability was assigned to both tails of the distribution. A high (low) box indicates that too much probability mass was assigned to low (high) outcomes. Although the box locations and sizes vary over individual, there is a preponderance of upper ends of the boxes above 0.75 for the Q1 survey output growth histograms: outcomes were 'too often' in the upper quartile of the forecast distributions. For the inflation forecasts, especially for the second, third and fourth quarter forecasts, there are a preponderance of boxes that fall short of 0.75 and extend below 0.25: too much weight assigned by forecasters to relatively high inflation rates. These findings hold irrespective of whether we fit generalized beta distributions and then read off the pit values, or calculate pits by linear interpolation.[17] The pit-based evaluation is consistent with the overall pessimism of the histogram forecasts noted in the Introduction, although we have arrived at this conclusion from an evaluation of the whole densities (by comparing them to the realized values) rather than a direct comparison of measures of central tendency from the histograms with point predictions (as in e.g., Engelberg *et al.* (2009)).

5 Calibration when the underlying cdf is observed

The approach suggested by Engelberg *et al.* (2009) is a flexible way of fitting continuous distribution functions to the reported histograms to enable the calculation of moments and pits. Moreover, the results are qualitatively unchanged if instead we fit gaussian distributions or suppose the probability mass is uniform within each histogram interval. However, it may be that none of these methods closely approximate the individuals' underlying subjective distributions. To determine whether the assumptions we have made result in the labelling of the histograms as being overly pessimistic, we assess the pits for the subset of the histograms which essentially completely reveal the individuals' underlying cdfs. To see this, note that the problem with the calculation of the pit outlined in (7) is that we do not know whether the probability mass in the interval $[3, 3.9)$ is close to 3, or to 4, or uniformly distributed as assumed in (7).[18] However, the histogram accurately reveals the pits for realized values equal to the 'x' values in table 1 - these are 'error-free' pits unaffected by any distributional assumptions. To obtain a meaningful sample of pits, we take those corresponding to

[17] Qualitatively similar results hold if we fit gaussian distributions: results not shown to save space.

[18] The two parametric distribution approaches essentially allow different assumptions about the distribution of probability mass in this interval: the location of the mass within this interval will depend on the probabilities assigned to the other intervals.

actual values which are 'close'[19] to the observed points on the cdf. As an example, suppose the realized value was 3.09 for the histogram forecast given in table 1. Using linear interpolation as in (7), we obtain $\Pr(Y < 3.09) = F(3) + \frac{3.09-3}{1}(F(4) - F(3)) = 0.5 + 0.09 \times 0.2$, and we suppose any error from the assumption that underlies the specific value we obtain for the second term ($0.09 \times 0.2 = 0.018$) is small.[20]

The set of 'boundary-actual' pits defined in this way is too small to permit the calculation of tests for each individual (and horizon) of the form reported in table 3. It is tempting to compare the distributions of the boundary and non-boundary pits across all individuals for first quarter surveys, and for second quarter surveys, etc. However the pits for any one survey will be correlated across individuals, so that the distributions of the pits across individuals and surveys will be unknown even for optimal forecasts. Instead, we consider the cross-sectional median pits. The empirical distributions of the median pits across surveys (for a given horizon) can be used to assess the accuracy with which the assumed distribution approximates the underlying subjective distribution.

Under the assumptions that *i*) the assumed distributions approximate the underlying subjective distributions and *ii*) the subjective distributions are approximately correctly calibrated, we would expect both the empirical distributions of the boundary and non-boundary median to be approximately $U(0,1)$. If the boundary pits are $U(0,1)$, but the non-boundary pits are not, then the econometrician is at fault: the generalized beta distribution, for example, does not do a good job of characterizing the individual-level subjective probability distributions. If the boundary pits themselves are not $U(0,1)$, then we conclude that the underlying distributions are poor approximations to the true predictive densities.

Figure 9 plots boundary and non-boundary pits separately for inflation, where the pits are calculated after fitting distributions as in Engelberg *et al.* (2009). For inflation there are a maximum of 10 boundary and 32 non-boundary pits, depending on the survey quarter,[21] so that the small samples on which these box plots are based needs to be borne in mind. For output growth there were too few boundary pits (5) to allow a meaningful comparison of the distributions of the

[19] Our definition of 'close' depends on the histogram interval widths, which are either 1 (as in the example) or 2 (for both inflation and output growth for the surveys 1981:3 to 1991:4). For interval widths of 1 (2), we consider realized values which are within 0.1 (0.2) of the upper interval limit as generating 'error-free' pits. This implies that on average one fifth of all realized values give rise to such pits.

[20] When we fit a parametric distribution to the histogram, the estimated pit will not exactly equal the cdf probabilities even for integer actuals unless the distribution fits the cdf points exactly (as when there are exactly three non-zero intervals in the case of the generalized beta distribution, for example).

[21] For first quarter surveys, for example, we have 42 sets of forecasts for the quarters from 1969:1 to 2010:1 (inclusive), but with 1985:1 and 1986:1 excluded as there were doubts about the period to which the forecasts made in these two quarters referred: see the SPF documentation.

two sets of pits. For inflation the evidence provided by the boundary pits is that the subjective probability distributions assign too much weight to high rates of inflation (relative to the realized outcomes). For correctly calibrated forecast densities the 'boxes' in figure 9 would span 0.25 to 0.75 on the vertical axis. Figure 10 uses linear interpolation of the histograms to calculate the pits. As expected, we obtain a similar picture for the boundary pits because both the fitting of the generalized beta distribution and linear interpolation are essentially equivalent for the upper histogram interval points. Consider now the non-boundary pits shown in figure 9 (figure 10 is essentially the same). The non-boundary pits are better calibrated. Taken at face value, this suggests that fitting generalized beta distributions distorts the individuals' subjective assessments, but in a way which offsets the pessimism that characterises those assessments.

Rather than considering the entire distributions of the boundary and non-boundary pits (as in figures 9 and 10), given the small numbers of observations we also formally tested simply whether the locations of the distributions differ. For each quarter of the year, we regress the annual time series of the cross-sectional median (of the z^*) on a constant, and test whether the constant is significantly different from zero, as a formal test of whether the pits are zero mean, as a necessary condition for the generalized beta distributions being correctly calibrated. We then run a further regression which includes a dummy variable taking the value 1 for boundary actual values, and test whether the coefficient on the dummy variable is significantly different from zero. If so, we conclude that the generalized beta distribution distorts the underlying probability assessments. We also report results for using the normal distribution in place of the beta distribution, and for using linear interpolation, as sensitivity checks.

Although there were too few boundary-pits for output growth to be able to consider differences in their empirical distribution relative to that of the non-boundary pits, we are able to test whether the mean of the pits differs significantly between the two sets.

The results are shown in table 4. Consider the results in the top panel for the beta distribution. For inflation, we reject the null that the mean is not significantly different from zero for surveys made in all but the first quarter of the year, inferring that the distributions underlying the pit calculations are not well calibrated. Further, the boundary dummies are clearly significant for all four quarters, signifying that the beta distribution distorts the underlying distributions. For output growth the first-quarter surveys are overly pessimistic (too much mass is assigned to low rates of output growth in the histogram that gives the median pit), but there is no evidence that the assumption of a beta distribution is at odds with the subjective distributions. The remaining

two panels show that essentially the same results hold for the two other distributional assumptions.

A possible explanation of our findings for inflation is that the occurrences of boundary-actual values of inflation just happen to be associated with lower than expected rates of inflation. The conclusion that the assumed parametric distributions poorly approximate the underlying subjective distributions requires that the outturns giving rise to the two sets of pits do not differ in a systematic fashion. There is no reason why this should be the case, but nevertheless it might be true of our historical sample. We can address this issue by estimating the average expected rates of inflation from (the median of) the survey respondents' point predictions of annual inflation. Table 5 reports the RMSE and bias of the median forecasts by survey quarter for each of the two sets of outturns. The RMSEs are roughly comparable, but the forecast biases of the boundary actuals are markedly negative for all four survey quarters, in contrast to the non-boundary actual forecast errors, indicating that boundary-actuals were associated with lower than expected inflation rates. For the first two survey quarters, on average the median expected forecast error was roughly minus a quarter of a percentage point for outturns 'close' to points on the cdf. A formal test of the equality of the population means (i.e., bias) of the two sets of forecasts rejects at the 5% level for all but the fourth quarter survey forecasts. Although in principle the boundary-actual pits enable a test of the underlying subjective distributions, in practice this is not possible for our sample because the boundary-actual values are associated with systematically lower than expected rates of inflation.

6 Conclusions

In recent times density forecasts have taken centre-stage in policy-based economic forecasting, especially for forecasting inflation. For example, every quarter the Bank of England publishes density forecasts of the annual rate of retail price inflation made by the UK Monetary Policy Committee (MPC), and many other central banks have followed suite. This break with the traditional concern of forecasting the central tendency or most likely outcome of the future value of the variable is understandable given that an assessment of the degree of uncertainty surrounding a point forecast is generally indispensable in a policy context.

The shift of focus to the forecast density has led some to question the value of survey respondents' point predictions, and the recommendation that only density forecasts be elicited. Based on our analysis of the point predictions and histograms of the respondents to the SPF, we question whether professional forecasters are sufficiently skilled at presenting their point predictions in the

form of histograms that the point predictions themselves can be dispensed with. We show that the updating of point predictions made by individual respondents can be reasonably well explained by forecasters revising their forecasts as new information becomes available, whereas estimates obtained from histograms often imply implausible parameter values for the underlying learning model. Formally, when we test that an individual applies the same weight to new information, and interprets that information in the same way, when he/she updates their point predictions and histogram means, then across the forecasters taken together we tend to reject for the short horizon forecasts.

We are careful to ensure that our results are not overly sensitive to the distributional assumptions we need to make. The individuals' probability assessments are reported as histograms, and further assumptions are required to obtain continuous distributions to allow the calculation of moments and probability integral transforms. We propose a way of assessing whether the distributional assumptions we make are appropriate, based on calculating probability integral transforms for the subset of observations for which the histograms are (nearly) fully informative without additional assumptions. Although in principle this enables an assessment of the underlying subjective distributions, in practice we find that this subset happens to be characterised by lower than expected rates of inflation. Nevertheless, our sensitivity checks suggest the results are robust across the different distributional assumptions we make. Even using 'best-practice' methods to fit distributions to the histograms we find that there is a case for continuing to elicit respondents' point predictions directly.

References

Atkeson, A., and Ohanian, L. (2001). Are Phillips Curves useful for forecasting inflation?. *Federal Reserve Bank of Minneapolis Quarterly Review*, **25**, 2–11. (1).

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

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business and Economic Statistics*, **19**, 465–474.

Boero, G., Smith, J., and Wallis, K. F. (2008a). Evaluating a three-dimensional panel of point forecasts: The Bank of England Survey of Economic Forecasters. *International Journal of Forecasting*, **24**, 354–367.

Boero, G., Smith, J., and Wallis, K. F. (2008b). Uncertainty and disagreement in economic prediction: the Bank of England Survey of Economic Forecasters. *Economic Journal*, **118**, 1107–1127.

Clements, M. P. (2009). Internal consistency of survey respondents' forecasts: Evidence based on the Survey of Professional Forecasters. In Castle, J. L., and Shephard, N. (eds.), *The Methodology and Practice of Econometrics. A Festschrift in Honour of David F. Hendry. Chapter 8*, pp. 206–226. Oxford: Oxford University Press.

Clements, M. P. (2010). Explanations of the Inconsistencies in Survey Respondents Forecasts. *European Economic Review*, **54**, 536–549.

Croushore, D. (1993). Introducing: The Survey of Professional Forecasters. *Federal Reserve Bank of Philadelphia Business Review*, **November/December**, 3–13.

Croushore, D., and Stark, T. (2001). A real-time data set for macroeconomists. *Journal of Econometrics*, **105**, 111–130.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). Evaluating density forecasts: With applications to financial risk management. *International Economic Review*, **39**, 863–883.

Doornik, J. A., and Hansen, H. (1994). A practical test for univariate and multivariate normality. Discussion paper, Nuffield College.

Engelberg, J., Manski, C. F., and Williams, J. (2009). Comparing the point predictions and subjective probability distributions of professional forecasters. *Journal of Business and Economic Statistics*, **27**, 30–41.

Fixler, D. J., and Grimm, B. T. (2005). Reliability of the NIPA estimates of U.S. economic activity. *Survey of Current Business*, **85**, 9–19.

Fixler, D. J., and Grimm, B. T. (2008). The reliability of the GDP and GDI estimates. *Survey of Current Business*, **88**, 16–32.

García, J. A., and Manzanares, A. (2007). Reporting biases and survey results: evidence from European professional forecasters. ECB Working Paper No. 836, European Central Bank, Frankfurt.

Giordani, P., and Söderlind, P. (2003). Inflation forecast uncertainty. *European Economic Review*, **74**, 1037–1060.

Hall, S. G., and Mitchell, J. (2009). Recent developments in density forecasting. In Mills, T. C., and Patterson, K. (eds.), *Palgrave Handbook of Econometrics, Volume 2: Applied Econometrics*, pp. 199–239: Palgrave MacMillan.

Kandel, E., and Zilberfarb, B. Z. (1999). Differential interpretation of information in inflation forecasts. *The Review of Economics and Statistics*, **81**, 217–226.

Lahiri, K., and Sheng, X. (2008). Evolution of forecast disagreement in a Bayesian learning model. *Journal of Econometrics*, **144**, 325–340.

Lahiri, K., and Sheng, X. (2010). Learning and heterogeneity in GDP and inflation forecasts. *International Journal of Forecasting*, **26**, 265–292.

Landefeld, J. S., Seskin, E. P., and Fraumeni, B. M. (2008). Taking the pulse of the economy. *Journal of Economic Perspectives*, **22**, 193–216.

Manzan, S. (2011). Differential interpretation in the Survey of Professional Forecasters. *Journal of Money, Credit and Banking*, **43**, 993–1017.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*: John Wiley and Sons, Ltd.

Patton, A. J., and Timmermann, A. (2010). Why do forecasters disagree? lessons from the term structure of cross-sectional dispersion. *Journal of Monetary Economics*, **57**, 803–820.

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, **74**, 967–1012.

Stock, J. H., and Watson, M. W. (2007). Why has U.S. Inflation Become Harder to Forecast?.

Journal of Money, Credit and Banking, **Supplement to Vol. 39**, 3–33.

Stock, J. H., and Watson, M. W. (2010). Modelling Inflation after the Crisis. *NBER Working Paper Series*, **16488**.

Tay, A. S., and Wallis, K. F. (2000). Density forecasting: A survey. *Journal of Forecasting*, **19**, 235–254. Reprinted in Clements, M. P. and Hendry, D. F. (eds.) *A Companion to Economic Forecasting*, pp.45 – 68, Oxford: Blackwells (2002).

Zarnowitz, V., and Braun, P. (1993). Twenty-two years of the NBER-ASA quarterly economic outlook surveys: aspects and comparisons of forecasting performance. In Stock, J., and Watson, M. (eds.), *Business Cycles, Indicators, and Forecasting*, pp. 11–84: Chicago: University of Chicago Press and NBER.

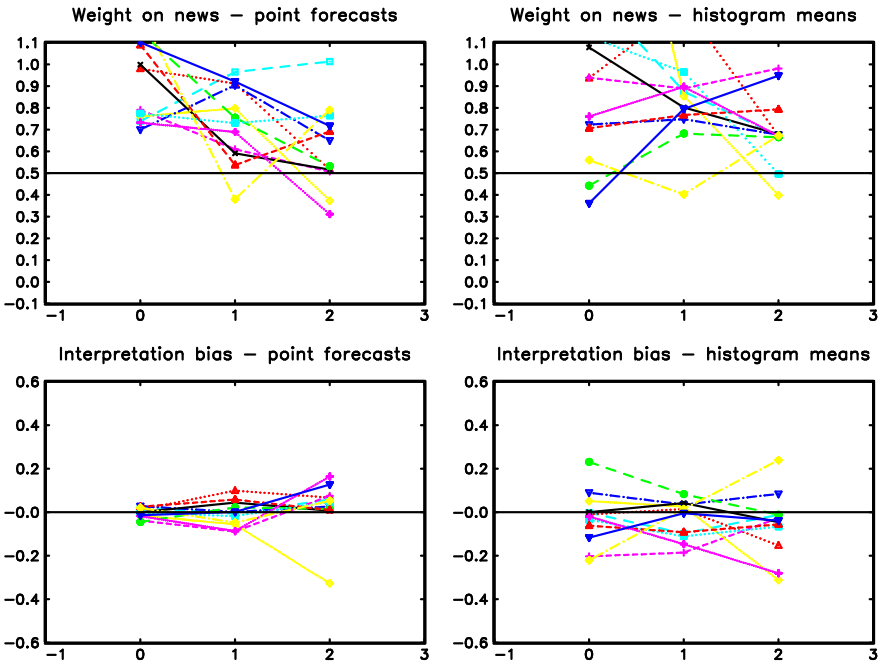

Figure 1: Output. Forecast as signal. Generalized beta.

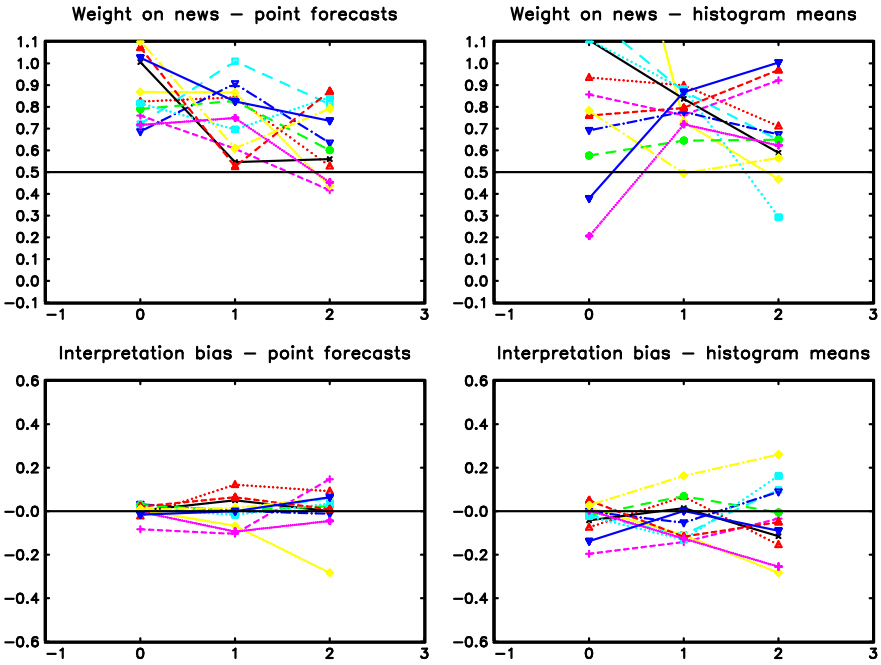

Figure 2: Output. Actual as signal. Generalized beta.

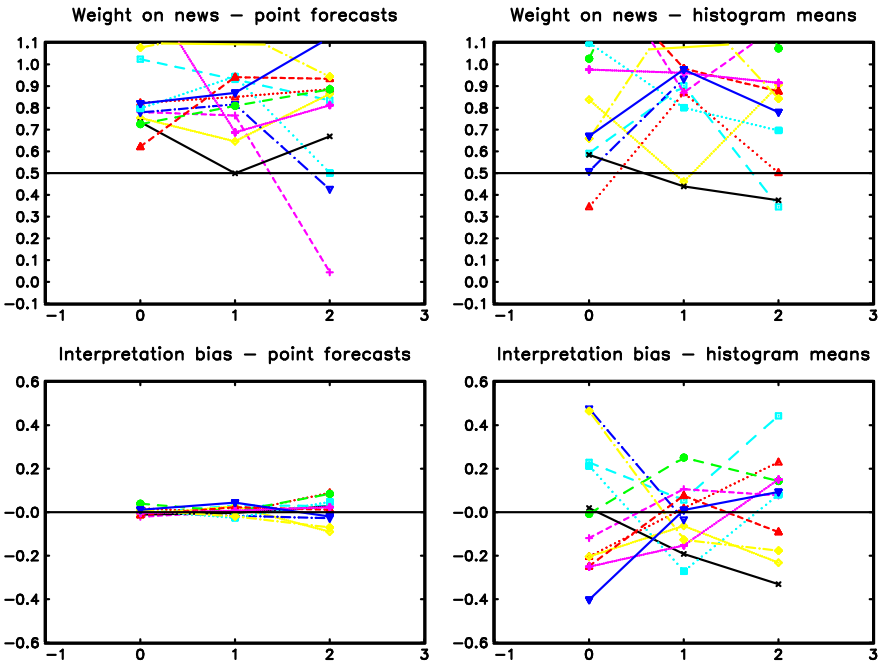

Figure 3: Inflation. Forecast as signal. Generalized beta.

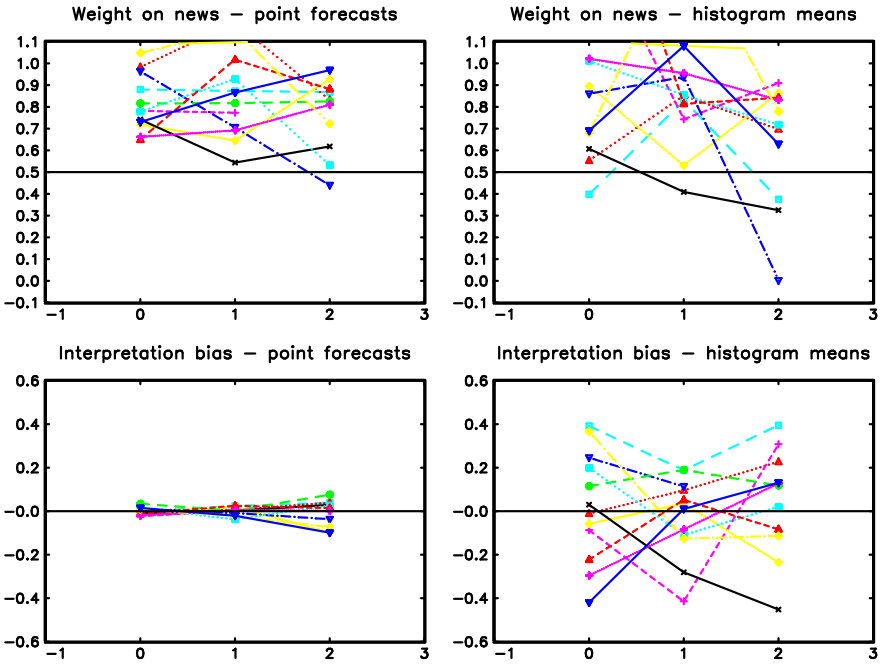

Figure 4: Inflation. Actual as signal. Generalized beta.

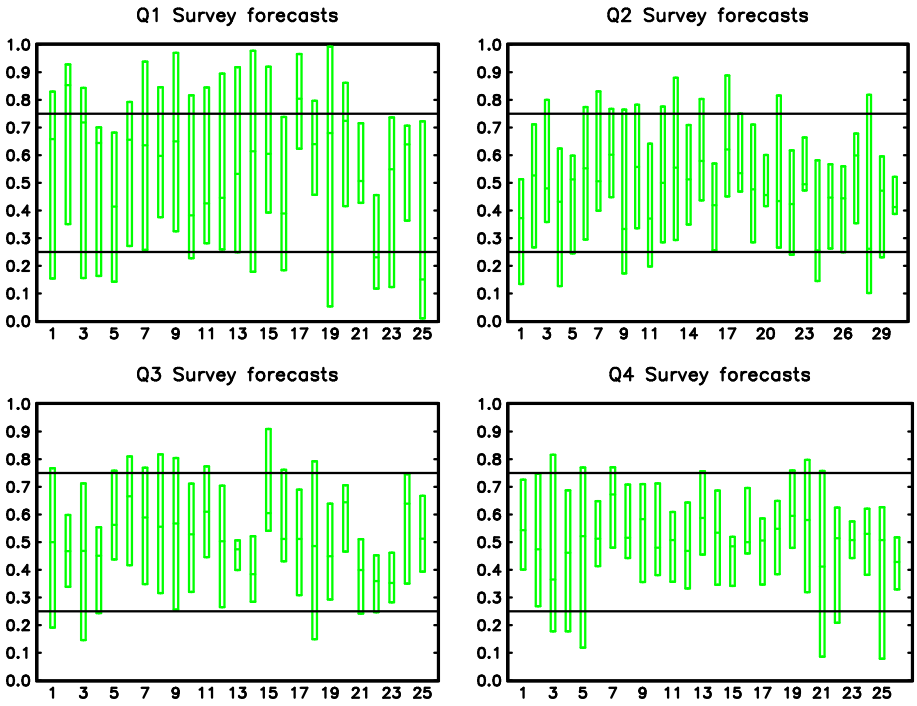

Figure 5: Output growth. Box plots of z's for each respondent with # z more than 10. The box is the interquartile range. z's calculated by fitting Generalized Beta distributions.

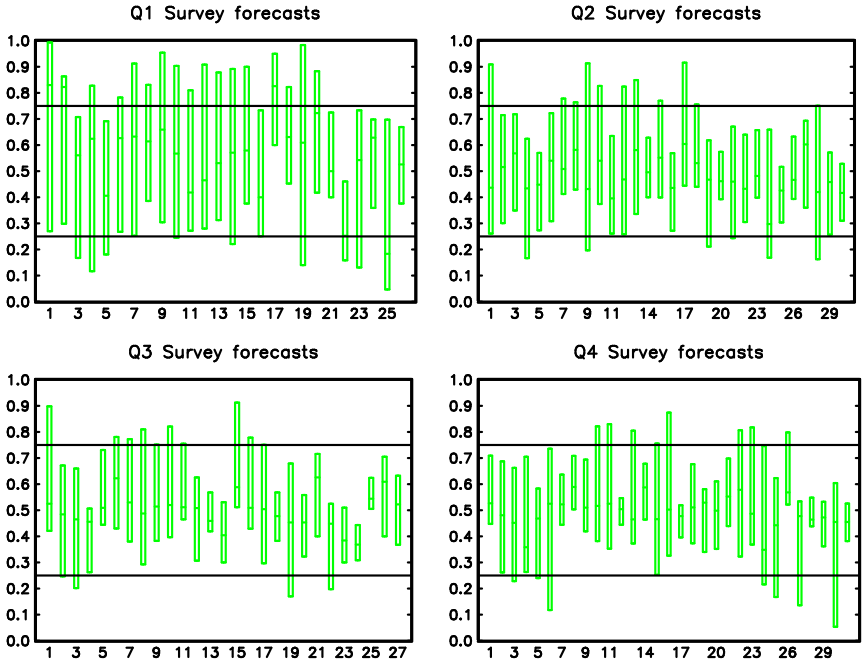

Figure 6: Output growth. Box plots of z's for each respondent with # z more than 10. The box is the interquartile range. z's calculated by linear interpolation.

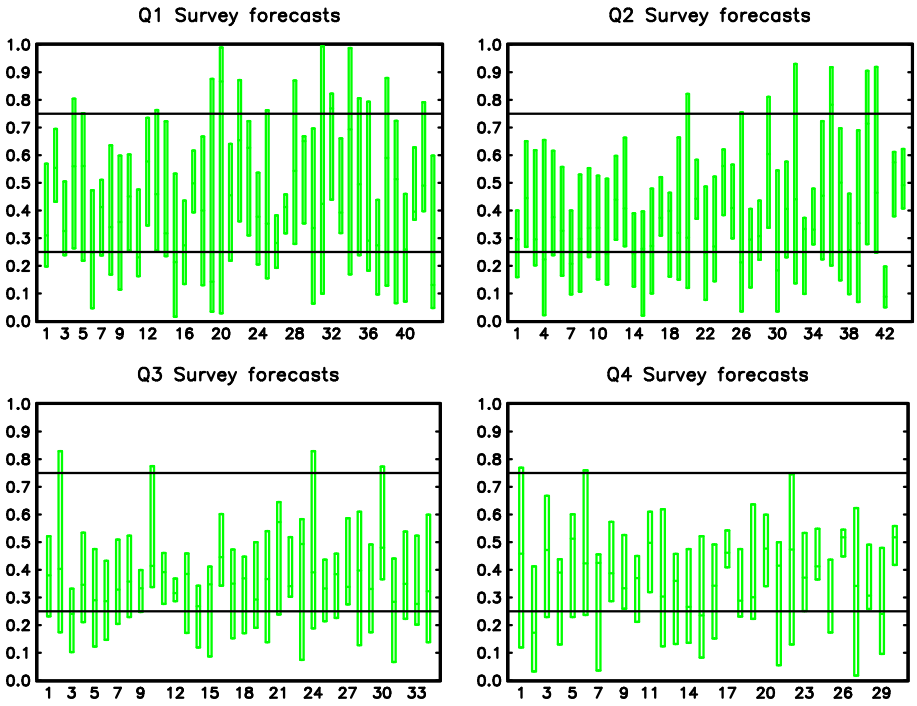

Figure 7: Inflation. Box plots of z's for each respondent with # z more than 10. The box is the interquartile range. z's calculated by fitting Generalized Beta distributions.

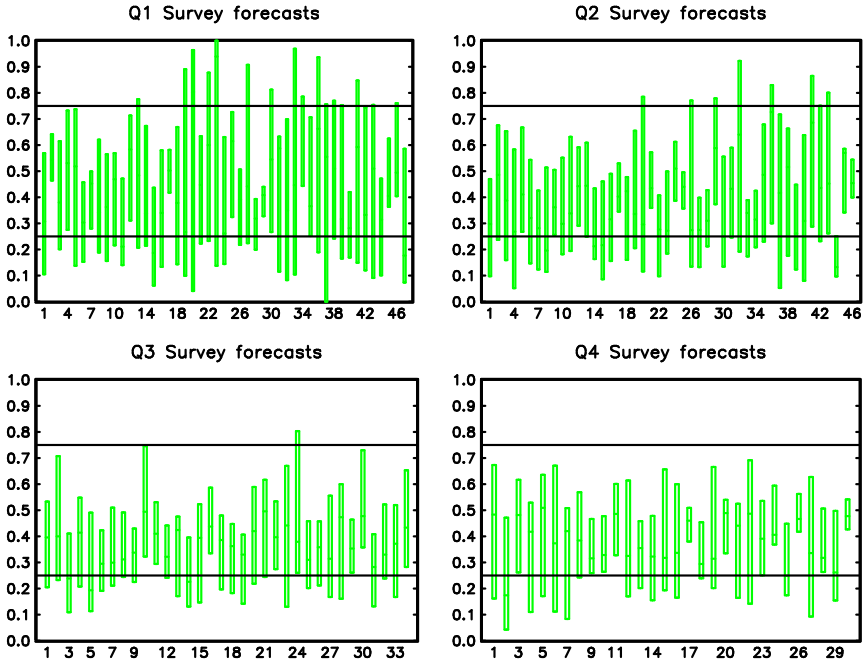

Figure 8: Inflation. Box plots of z's for each respondent with # z more than 10. The box is the interquartile range. z's calculated by linear interpolation.

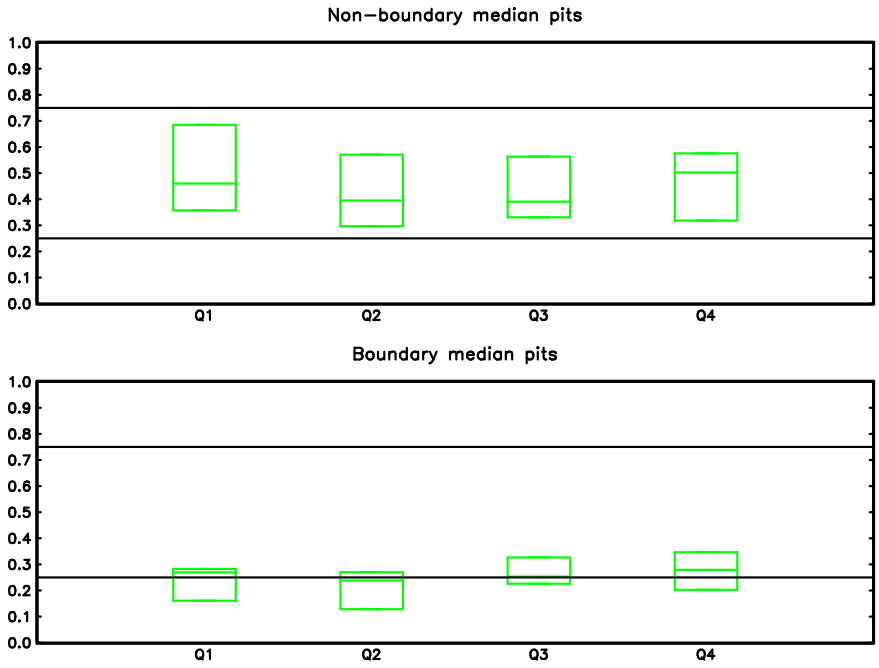

Figure 9: Inflation. Median pit. Generalised beta distribution.

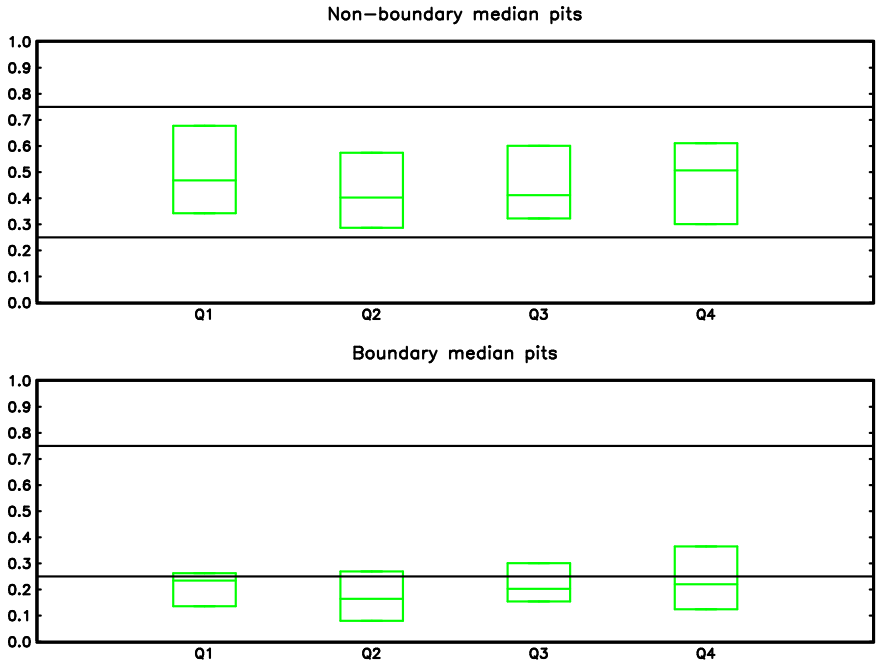

Figure 10: Inflation. Median pit. Linear interpolation of the histogram.

Table 1: Example of a histogram return.

Interval	Probability	x	$F(x)$
$`< -2'$	0.0	-2	0
-2 to -1.1	0.0	-1	0
-1 to -0.1	0.0	0	0
0 to 0.9	0.0	1	0
1 to 1.9	0.0	2	0
2 to 2.9	0.50	3	0.5
3 to 3.9	0.20	4	0.7
4 to 4.9	0.30	5	1.0
5 to 5.9	0.0	6	1.0
$`6+'$	0.0	7	1.0

Table 2: Tests that the BLM parameters for the point predictions and histogram means are equal for each respondent

		$h=2$				$h=1$				$h=0$			
	L_{th}	T_i	Ave T_{n_i}	$\alpha_i^{(1)}=\alpha_i^{(2)}$	$\mu_i^{(1)}=\mu_i^{(2)}$	T_i	Ave T_{n_i}	$\alpha_i^{(1)}=\alpha_i^{(2)}$	$\mu_i^{(1)}=\mu_i^{(2)}$	T_i	Ave T_{n_i}	$\alpha_i^{(1)}=\alpha_i^{(2)}$	$\mu_i^{(1)}=\mu_i^{(2)}$
Fitting Generalized Beta distribution: Engelberg *et al.* (2009) approach													
Inflation	F	28	13	0.30	0.35	21	14	0.58	0.43	17	14	0.00	0.33
Inflation	A	28	13	0.16	0.39	21	14	0.47	0.24	17	14	0.00	0.09
Output	F	20	13	0.45	0.21	21	13	0.01	0.23	18	13	0.00	0.69
Output	A	20	13	0.59	0.27	21	13	0.02	0.00	18	13	0.00	0.91
Uniform probability mass													
Inflation	F	28	13	0.41	0.28	22	14	0.17	0.61	18	14	0.00	0.44
Inflation	A	28	13	0.27	0.43	22	14	0.40	0.50	18	14	0.00	0.10
Output	F	20	13	0.59	0.16	21	13	0.02	0.16	18	13	0.00	0.71
Output	A	20	13	0.64	0.27	21	13	0.02	0.00	18	13	0.00	0.98
Normal approximation to histogram forecasts													
Inflation	F	28	13	0.69	0.25	21	13	0.72	0.48	17	14	0.00	0.14
Inflation	A	28	13	0.43	0.21	21	13	0.41	0.07	17	14	0.00	0.04
Output	F	20	13	0.13	0.74	21	13	0.01	0.18	18	13	0.00	0.85
Output	A	20	13	0.51	0.61	21	13	0.16	0.23	18	13	0.00	0.84

L_{th} = F implies the signal is the forecast value; = A implies the latest available. T_i is the number of respondents, and 'Ave T_{n_i}' indicates the average number of forecasts per respondent. The table reports p-values of the null hypotheses that $\alpha_i^{(1)}=\alpha_i^{(2)}$ and $\mu_i^{(1)}=\mu_i^{(2)}$ where in both cases $i=1,2,\ldots,T_i$.

Table 3: Proportion of individuals for whom we reject the null that their histograms are 'accurate' based on probability integral transform tests

	$H_0 : z^*$ normal	$H_0 : z^*$ IID(0,1)	$H_0 : z^*$ normal	$H_0 : z^*$ IID(0,1)
	z^*'s calculated by fitting Generalized Beta distributions			
Survey	Inflation		Output growth	
1	0.24	0.48	0.13	0.75
2	0.21	0.67	0.40	0.25
3	0.33	0.53	0.46	0.31
4	0.44	0.56	0.67	0.33
	z^*'s calculated by linear interpolation of the histograms			
	Inflation		Output growth	
1	0.22	0.37	0.00	0.47
2	0.15	0.46	0.22	0.26
3	0.06	0.59	0.38	0.19
4	0.26	0.42	0.60	0.15
	z^*'s calculated by a normal approximation to the histograms			
	Inflation		Output growth	
1	0.44	0.70	0.05	0.68
2	0.38	0.81	0.32	0.55
3	0.53	0.65	0.56	0.50
4	0.63	0.53	0.70	0.70

For each individual who reported more than 10 histogram forecasts of a given horizon, we calculate a test of the normality of their $\{z_t^*\}$ (headed $H_0 : z^*$ normal) and a three-degree of freedom likelihood ratio test of zero-mean, unit variance and independence (specifically, zero first-order autocorrelation) using gaussian likelihood functions (headed $H_0 : z^*$ IID(0,1)). The entries in the table are the proportion of individuals for which we reject the null at the 10% level. For each survey quarter there are generally around 15 forecasters who have responded to 10 or more surveys. The left column denotes the survey quarter, where survey quarter '1', for example, corresponds to the longest horizon forecasts.

The three panels relate to three different ways of calculating the probability integral transforms from the histograms. Technical note: z's calculated as 0 or 1 (because the realization lies outside the intervals with non-zero weights, when calculated by linear interplation assuming probability mass is uniformly distributed within a interval) are replaced by 0.01 and 0.99 so that the corresponding z_t^* is defined.

Table 4: Analysis of median and cross-sectional dispersion of pits at 'boundary actuals'

Generalized Beta distributions				
	Inflation			
	δ_0	p-value $\delta_0 = 0$	α_1	p-value $\alpha_1 = 0$
1	-0.06	0.69	-0.78	0.00
2	-0.24	0.08	-0.87	0.00
3	-0.19	0.02	-0.52	0.00
4	-0.19	0.05	-0.49	0.00
	Output growth			
	δ_0	p-value $\delta_0 = 0$	α_1	p-value $\alpha_1 = 0$
1	0.49	0.04	0.55	0.48
2	0.09	0.45	0.19	0.57
3	0.07	0.45	-0.05	0.86
4	-0.02	0.91	-0.16	0.82
Linear interpolation				
	Inflation			
	δ_0	p-value $\delta_0 = 0$	α_1	p-value $\alpha_1 = 0$
1	-0.10	0.50	-0.89	0.00
2	-0.27	0.03	-0.83	0.00
3	-0.23	0.01	-0.60	0.00
4	-0.24	0.04	-0.66	0.00
	Output growth			
	δ_0	p-value $\delta_0 = 0$	α_1	p-value $\alpha_1 = 0$
1	0.34	0.08	0.38	0.56
2	0.08	0.54	0.32	0.41
3	0.09	0.43	0.11	0.77
4	0.05	0.74	0.07	0.91
Normal approximation				
	Inflation			
	δ_0	p-value $\delta_0 = 0$	α_1	p-value $\alpha_1 = 0$
1	0.02	0.90	-0.95	0.00
2	-0.28	0.05	-0.82	0.00
3	-0.27	0.02	-0.62	0.00
4	-0.20	0.15	-0.74	0.00
	Output growth			
	δ_0	p-value $\delta_0 = 0$	α_1	p-value $\alpha_1 = 0$
1	0.38	0.09	0.31	0.65
2	0.05	0.76	0.29	0.51
3	0.12	0.37	0.03	0.95
4	0.06	0.78	-0.39	0.68

The entries in the table are the estimates of δ_0 in the regression of $x = \delta_0 + \varepsilon_t$, along with the p-value that $\delta_0 = 0$, as well as the estimates of α_1 in $x = \alpha_0 + \alpha_1 D_t + \varepsilon_t$, along with the p-value that $\alpha_1 = 0$, all using HCSEs. x is the median pit (z^*), and D_t defines boundary-actual pits. That is, D_t is a dummy variable that takes the value 1 when the actual value (used in the pit calculation) is close to a histogram interval boundary.

Table 5: Properties of the median expected forecast errors for the boundary and non-boundary actual observations

Survey qtr.	RMSFE		Forecast bias		p-value of test of equal bias
	boundary actual	non-bondary	boundary actual	non-bondary	
1	0.59	0.89	-0.27	0.28	0.024
2	0.35	0.54	-0.28	0.09	0.004
3	0.21	0.32	-0.17	-0.02	0.036
4	0.13	0.18	-0.08	0.02	0.051

The last column is the p-value of a t-test that the population means of the two sets of forecast errors (those corresponding to outturns close to cdf points - boundary actuals, and those corresponding to actuals not close to the boundary) are equal when we allow for unknown and unequal variances. The test requires the populations are normally distributed or the sample sizes are large.

Appendix. Illustrative calculations for a survey respondent

An individual's inflation histograms in response to the 2000:1 and 2000:2 surveys are given in table 6. Using our preferred method of obtaining means from histograms, we estimate the means as 1.69 and 4.00. (See Engelberg *et al.* (2009), p.37 for fitting methods when there are two non-zero intervals). So between the first and second quarter of the year, the mean rate of annual year-on-year inflation for 2000 over 1999 more than doubles. Throughout we have emphasized errors due to the histograms only imperfectly revealing the underlying subjective distributions, whereby the assumption that the histograms can be approximated by generalized beta distributions may be at odds with the underlying subjective probability distributions. To guard against elicitation errors, following Engelberg *et al.* (2009) we calculate the extreme values that the mean could take: these are lower and upper bounds of 1.15 and 2.15 for 2000:1, and of 3.5 and 4.5 for 2000:2.[22]

So the smallest possible increase in the mean is from 2.15 to 3.5 (2000:1 upper bound to 2000:2 lower bound). There has clearly been a marked increase in the forecast inflation rate implied by the histograms. This is entirely plausible, but becomes somewhat suspect when taken together with the point predictions for the annual rate of inflation by this respondent to the same two surveys. In their 2000:1 survey return, forecasts are given of the price level for the 4 quarters of the year, as well as the annual level. These are consistent for this survey return: the average of the quarters equals the annual. Given the latest estimate (at the time of the survey) of the annual level for 1999 of 104.32, the point prediction of the growth rate made in the first quarter of the year is 1.74 - close to the histogram mean, and within the bounds on the mean. The 2000:2 survey return also contains a forecast of the annual level (which is again consistent with the forecasts of the levels for each quarter). Last year's level has been revised up a little, to 104.55. The forecast of the annual level is a little higher than last quarter, resulting in an annual inflation rate forecast of just over 2 (2.02). This is outside the bounds on the histogram mean, and an instance of an inconsistent pair of forecasts as documented by Engelberg *et al.* (2009).

Given that the latest actual quarterly inflation rate at the time the return was made to the second quarter survey was just 1.5 (at an annual rate, corresponding to the first quarter of the year), as well as the point forecasts, the increase in the forecast mean implied by the 2000:2 histogram appears unwarranted. The point forecasts may better represent this respondent's inflation outlook.

[22] These bounds on the mean require only that the histogram upper limits constitute points on the individual's cdf.

Table 6: A respondent's 2000:1 and 2000:2 histogram forecasts of annual inflation.

	2000:1 Survey	2000:2 Survey
$`< -0'$	0	0
0 to 0.9	0	0
1 to 1.9	85	0
2 to 2.9	15	0
3 to 3.9	0	50
4 to 4.9	0	50
5 to 5.9	0	0
6 to 6.9	0	0
7 to 7.9	0	0
$`8+'$	0	0