econometrics

MDPI

Article

Are Some Forecasters' Probability Assessments of Macro Variables Better than those of Others?

Michael P. Clements

ICMA Centre, Henley Business School, University of Reading, Reading RG6 6BA, UK; m.p.clements@reading.ac.uk

Received: 24 November 2018; Accepted: 2 May 2020; Published: 6 May 2020

Abstract: We apply a bootstrap test to determine whether some forecasters are able to make superior probability assessments to others. In contrast to some findings in the literature for point predictions, there is evidence that some individuals really are better than others. The testing procedure controls for the different economic conditions the forecasters may face, given that each individual responds to only a subset of the surveys. One possible explanation for the different findings for point predictions and histograms is explored: that newcomers may make less accurate histogram forecasts than experienced respondents given the greater complexity of the task.

Keywords: survey forecasters; probability distributions; probability scores

PACS: C53; E37

1. Introduction

A natural question to ask of a group of forecasters is whether some are better than others, or is the fact that forecaster A makes a more accurate forecast than B this period simply due to chance, so that the situation is as likely as not to be reversed next time? In the context of forecasting inflation and GDP growth one quarter and one year ahead, D'Agostino et al. (2012) find little evidence to suggest that some forecasters are 'really better'. In our paper, we apply their methodology to the probability assessments of GDP growth and inflation reported to the US Survey of Professional Forecasters (SPF). There have been a number of assessments of the individuals' probability assessments, but none that directly address the question of interest: are some forecasters' probability assessments more accurate than those of other forecasters? For example, Clements (2018) compares the individual forecasts to a benchmark, but does not consider whether there are systematic differences between forecasters.

Why does it matter whether some forecasters are better than others? In recent years, there has been much work on expectations formation. If some forecasters were superior to others, this might have a bearing on some of the ingredients of the models of expectations. For example, some of the models currently discussed in the literature replace the assumption of full-information rational expectations—in which all agents know the true structure of the economy and have access to the same information set—with 'informational rigidities'. Agents are assumed to form their expectations rationally subject to the information constraints they face. For example, under noisy information agents only ever observe noisy signals about economic fundamentals.[1] The baseline noisy information model assumes the noise-variance contaminating agents' signals is equal across agents. Consequently, agents' forecasts are all equally as good. If this is false, then the homogeneous signal assumption

[1] See, e.g., (Woodford 2002) and (Sims 2003).

could be dropped, although Coibion and Gorodnichenko (2012, 2015) argue the macro-level evidence supports the baseline model.[2]

D'Agostino et al. (2012) conclude that there are no real differences between forecasters in terms of the accuracy of their point forecasts. Why then might we expect differences in terms of the accuracy of their probability assessments? Producing histogram forecasts is typically more time-consuming and costly than producing point predictions, for a number of reasons. Point predictions are often reported in the media, and made available in reports that professional forecasters are likely to have ready access to. Moreover, professional forecasters may need to produce point forecasts regularly, and not just at the behest of the SPF. There is less likely to be a prevailing view about the probability assessment, that respondents are able to draw on, and respondents may only produce such assessments for their SPF survey returns. There is then a more uneven playing field, and more scope for some forecasters to outperform others.

In addition, there is evidence in the literature that the probability assessments and point predictions do not always tally in terms of what they imply about the most likely outcome, or the expected outcome: see (Engelberg et al. 2009) and (Clements 2009, 2010, 2014b) for the US SPF. Hence there should be no presumption that the findings for point predictions carry over to the histograms.

Point forecasts are, of course, less valuable than probability density forecasts, because they do not provide information on the range of likely outcomes. This point has been made repeatedly in the literature. For example, in discussing point forecasts, (Granger and Newbold 1986, p. 149) state 'It would often be very much better if a more sophisticated type of forecast were available, such as a confidence interval with the property that the probability that the true value would fall into this interval takes some specified value.' See also (Chatfield 1993) for a discussion of interval forecasts. Some surveys such as the US SPF provide the 'sophisticated' forecasts Granger and Newbold wished for. Clements (2019b) provide a general treatment of survey expectations, and Castle et al. (2019) discuss forecast uncertainty for a non-technical audience. The focus of this paper—whether there are real differences in terms of the accuracy of individuals' probability assessments—is at least as important as asking this question of point forecasts.

The plan of the remainder of the paper is as follows. Section 2 describes the SPF forecast data. Section 3 describes the bootstrap test of D'Agostino et al. (2012), applied by those authors to the US SPF point predictions, and section 4 our application of their test to the SPF probability assessments. Section 5 presents the results. Section 6 compares our findings to those from using alternative approaches that have been applied to point predictions. Section 7 offers some concluding remarks.

2. Forecast Data: SPF Respondents' Forecasts

The US Survey of Professional Forecasters (SPF) is a quarterly survey of macroeconomic forecasters of the US economy that began in 1968, and is currently administered by the Philadelphia Fed (see Croushore 1993). The SPF is made freely available by the Philadelphia Fed. It is perhaps the foremost survey of its kind, and a staple for academic research: an academic bibliography[3] listed 101 papers as of February 2018.

We use the 150 surveys from 1981:3 to 2018:4 inclusive, and consider the probability assessments of output growth (real GDP) and GDP deflator inflation, as well as the point predictions for these two variables.

The survey asks for histogram forecasts of the annual rates of growth (of GDP, and the GDP deflator) for the year of the survey relative to the previous year, as well as of the next year, relative to the current year. This results in a sequence of 'fixed-event' histogram forecasts. Using annual inflation

---

[2] See (Clements 2019a) for further discussion.

[3] http://www.phil.frb.org/research-and-data/real-time-center/survey-of-professional-forecasters/academic-bibliography.cfm.

in 2016 (compared to 2015) as an illustration, the 2016:Q1 survey yields a forecast of just under a year ahead, and the 2016:Q4 survey, a forecast with a horizon of just under a quarter. (We do not use the longer-horizon forecasts of 2016 which are provided in the 2015 surveys). This means we only have one histogram a year if we want a series of fixed-horizon forecasts: we could take the forecasts made in the first quarter, say, to give an annual series of year-ahead forecasts, or in the fourth quarter, to give an annual series of one-quarter-ahead forecasts.

As a result, there are only approximately one quarter as many histogram forecasts (of a given horizon) as there are point predictions used by D'Agostino et al. (2012).[4] Each survey also provides point predictions of quarterly growth rates for the current quarter and the next four quarters, providing a quarterly series of fixed-horizon forecasts—these are the forecasts used by D'Agostino et al. (2012). Although our main interest is in the probability assessments, we also apply the bootstrap test to a matching sequence of annual fixed-event point predictions. The SPF elicits forecasts each quarter of the annual calendar-year rates of growth of GDP and of inflation, matching the histogram forecasts. This serves as a check of whether the results for our annual series of point predictions are similar to those of D'Agostino et al. (2012) for the quarterly series, given that we will (of necessity) be working with an annual series for histograms.[5]

3. The Bootstrap Test

The test compares the empirical distribution of forecaster performance to the distribution which would be obtained under the null hypothesis of equal ability, constructed by randomly assigning forecasts to forecasters. The test is detailed in (D'Agostino et al. 2012, p. 718), and described below. It accounts for the unbalanced nature of the panel, and the possibility that comparisons across forecasters might be distorted by some forecasters being active at times which are conducive to more accurate forecasting, and others participating during more uncertain times.[6]

The changing composition of the SPF panel is evident from Table 1, which provides an illustrative snapshot of the (standardized) RPS statistics of the ten forecasters who filed the most inflation histogram responses to Q4 surveys over our sample period. (The ranked probability score (RPS)—which is based on the forecaster's histograms—and the meaning of 'standardized' are described subsequently). The panel changes because of entry and exit, as well as non-participation by members in particular periods (as documented by Engelberg et al. 2011, amongst others). Entry, exit and non-participation is clearly evident in Table 1. For example, the first forecaster joined the survey in 1991:Q4, and then remained an active respondent over the sample period, albeit with occasional periods of non-response. Forecaster 2 joined a year earlier. Forecaster 3 was active from the beginning of the sample, and left in 2009. There are of course more missing values for the less prolific SPF respondents.

---

[4] There are ways of constructing a quarterly series. For example, D'Amico and Orphanides (2008) construct an approximate year-ahead horizon rolling-event sequence of forecasts from the first-quarter current-year forecast, the second-quarter current-year forecast, the third-quarter next-year forecast, and the fourth-quarter next-year forecast. This gives a quarterly-frequency series of forecasts, but from the first to the fourth quarter the actual horizons are 4, 3, 6 and 5, and the target moves from the current to next year growth rate between the second and third quarters. It seems preferable to weigh the current and next year's forecasts, where the weights vary with the quarter of the year of the survey, reflecting the distances of the forecasts from the desired forecast horizon (see, e.g., D'Amico and Orphanides 2014). Knüppel and Vladu (2016) consider approximating rolling forecasts from fixed-event forecasts. Rather than use such approximations, we use the shorter sample of actual histogram forecasts.

[5] Note that there are subtler differences too, besides the smaller number of forecast observations, when we consider an annual calendar-year growth rate forecasts compared to quarterly growth rates. Annual growth rates are of course smoother, and the so-called 'carry-over effect' means that forecast uncertainty does not decline in the forecast horizon in the way which one might expect (see Tödter 2010; Clements 2014a).

[6] It is well-known that the difficulty of forecasting changes over time as macroeconomic conditions change, both over the long term (e.g., the Great Moderation: see McConnell and Perez-Quiros 2000) and due to short-term fluctuation, as evidenced by the use of models with time-varying heteroscedasticity (see, e.g., Clark 2011).

The unbalanced nature of panels of survey forecasters such as the US SPF is an impediment to inter-forecaster comparisons, but is allowed for in the approach of (D'Agostino et al. 2012, p. 718).[7]

Table 1. Illustration. Standardized anked probability score (RPS) statistic for Q4 survey inflation forecasts: the 10 forecasters with the most responses.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1981.4	.	.	0.00	.	.	.	.	.	2.22	.
1982.4	.	.	2.37	.	.	.	.	.	.	.
1983.4	.	.	0.02	.	.	.	.	.	.	.
1984.4	.	.	0.00	.	.	.	.	.	6.54	.
1985.4	.	.	0.00	.	.	.	.	.	2.36	.
1986.4	.	.	0.00	.	.	.	.	0.00	.	.
1987.4	.	.	0.00	.	.	.	.	.	0.97	.
1988.4	.	.	0.00	.	.	.	.	.	.	.
1989.4	.	.	0.01	.	.	.	.	0.04	.	.
1990.4	.	1.77	0.00	.	.	2.04	.	0.00	9.05	.
1991.4	2.85	0.71	0.00	0.71	0.93	0.76	.	.	2.85	.
1992.4	0.78	1.06	0.00	3.12	0.00	0.37	.	.	0.35	.
1993.4	4.23	0.02	0.00	.	0.00	0.62	0.01	9.92	0.20	.
1994.4	0.36	2.58	0.00	0.45	0.00	0.00	0.70	.	0.45	.
1995.4	1.38	0.59	1.65	1.52	0.81	1.35	0.30	.	1.05	0.41
1996.4	0.24	3.02	0.76	0.48	4.36	0.60	1.36	.	.	0.60
1997.4	0.08	0.00	1.35	0.69	2.76	0.11	.	1.38	2.76	0.00
1998.4	0.33	0.00	0.00	0.31	0.08	.	0.38	1.91	0.69	0.48
1999.4	.	0.42	0.00	0.21	0.00	.	0.84	0.00	0.42	0.42
2000.4	0.26	2.62	0.66	0.42	0.03	0.34	0.03	0.00	0.42	.
2001.4	2.13	.	0.00	0.80	0.02	4.24	0.33	0.07	0.26	.
2002.4	0.28	1.82	0.00	1.02	0.03	.	0.11	0.03	1.02	.
2003.4	0.16	3.46	0.00	0.32	0.13	.	.	0.00	0.51	3.75
2004.4	1.59	2.12	6.22	2.07	0.02	1.65	0.77	0.00	.	2.59
2005.4	0.13	1.10	0.01	.	0.00	0.03	0.67	0.00	0.82	5.01
2006.4	0.91	0.88	0.73	.	.	0.45	1.02	0.73	0.06	1.57
2007.4	.	1.83	0.00	.	.	1.47	0.04	.	0.65	0.27
2008.4	2.29	1.88	0.49	.	0.00	0.03	0.06	.	.	0.87
2009.4	0.56	2.69	0.00	.	.	0.43	0.01	.	.	0.86
2010.4	0.45	4.12	.	.	.	0.07	0.00	.	.	1.19
2011.4	1.36	1.27	.	0.06	2.16	0.02	0.04	.	.	.
2012.4	1.84	.	.	0.00	0.12	0.38	1.29	.	.	2.04
2013.4	0.24	3.01	.	0.24	0.00	.	0.48	.	.	.
2014.4	0.56	3.18	.	1.54	0.09	.	0.31	.	.	1.12
2015.4	0.69	1.31	.	0.35	0.00	.	.	.	.	0.79
2016.4	2.23	2.71	.	.	0.06	.	0.57	.	.	0.82
2017.4	.	3.15	.	.	0.01	.	0.46	.	.	0.58
2018.4	2.44	2.07	.	.	0.37	.	.	.	.	0.40

By comparing the empirical distribution with the null distribution, we can determine whether the 'best forecaster' occupies that position by virtue of being inherently the most accurate, or by chance, as well as whether the same is true at any percentile of the empirical distribution. For example, is the forecaster at the 25th percentile of the actual distribution inherently more accurate than three-quarters of the forecasters, or might this be due to chance? We use the bootstrap described below to calculate confidence intervals for the percentiles of the actual distribution, and also to calculate the probability of obtaining a lower (better) score by chance at a given percentile than that actually found.

[7] A possible issue with the approach might arise if a given forecaster's survey responses are serially dependent, because the bootstrap does not preserve intertemporal dependence. The scope of such dependencies, and the impact on the properties of the bootstrap test, are unclear.

The bootstrap test is as follows. Let s_{it} denote the 'score' for individual i in response to survey t, where we abstract from the forecast horizon. The score is either the squared forecast error, when we consider point forecasts, or the QPS or RPS value (as given by (1) and (2), and discussed below), when we consider the histogram forecasts. In either case, the score is non-negative, and zero indicates a perfect forecast. To account for different macroeconomic conditions, the scores are normalized by the average of all respondents' scores for that period (i.e., the scores for all the forecasts made at survey t):

$$S_{it} = s_{it} \left(\frac{1}{N_t} \sum_{t=1}^{N_t} s_{it} \right)^{-1},$$

where N_t is the number of such forecasts. S_{it} measures the performance of forecaster i relative to that of all others forecasting at that time. The overall mean score of forecaster i is then given by S_i:

$$S_i = \frac{1}{n_i} \sum_{t \in N_i} S_{it},$$

where N_i is the set of surveys to which i filed a response, and n_i is the number of elements in the set. Note inter-forecaster comparisons should be legitimate when forecasters i and j respond to different surveys and to different numbers of surveys, due to the normalization and averaging.[8]

For each survey t, the $S_{i\tau}$ (for $\tau = t$) are randomly assigned (with replacement) across a set of 'imaginary' forecasters who match the SPF forecasters exactly in terms of participation. For example, if the third SPF forecaster did not participate at time t, the third imaginary forecaster's time t forecast will also be missing. We continue for each t, for $t = 1, \ldots, T$. As stressed by D'Agostino et al. (2012), forecasters at each t can only be assigned a forecast from another forecaster made at that time.[9] The scores are then calculated for this random reshuffling, by averaging over the non-missing values for each i, to give the simulated vector of values S^1, with typical element S_i^1, where $i = 1, \ldots, N_f$, with N_f the number of actual forecasters. We repeat the above another 999 times to obtain 1000 bootstrap distributions, $\{S^1, \ldots, S^{1000}\}$.

Following D'Agostino et al. (2012), we compare selected percentiles of the actual distribution (e.g., the 'best', the forecaster occupying the position of the 5th percentile, etc.) to the 5th and 95th percentiles for those 'positions' calculated from the bootstrap distribution (generated under the assumption of equal accuracy). For example, to obtain a confidence interval for the best actual score under the null of equal forecast accuracy, we first calculate the best (minimum) score for each vector S^j, $j = 1, \ldots, 1000$. Denoting these by $\left\{ S^j_{\min} \right\}$, $j = 1, \ldots, 1000$, we then calculate the 50th and 950th largest values. If the best actual score lies within the confidence interval calculated under the null that forecasters are equally accurate, we do not reject the null at the 10% level. In addition, we calculate a 'p-value' as the proportion of the $\left\{ S^j_{\min} \right\}$ which are less than the actual best (say, $S_{\min}$). As shown below when we explain the results, this may provide useful additional information.

As another example, for the 5th percentile of the actual distribution, we select the 5th percentile values from each vector S^j, $j = 1, \ldots, 1000$, and then proceed as for the best, i.e., we calculate the 50th and 950th largest of the 5th-percentile-values, as well as the p-value.

---

[8] This may not be true when either forecaster responds to a small number of surveys. For example, if $n_i = 1$, and forecaster i happens to 'get lucky', with $s_{it} = 0$, forecaster i's performance will be artificially boosted (relative to what would be expected if $n_i > 1$). For this reason, we only consider forecasters who make a minimum of 5 forecasts at the horizon of interest.

[9] Our first imaginary forecaster '1' will have a participation pattern matching that of the first actual forecaster in Table 1. The forecaster will first be randomly assigned a score in 1991:Q4, and this will be drawn from the pool of 1991:Q4 actual standardized scores. The first imaginary forecaster will then feature in subsequent rounds of the survey, except in 1999:Q4, 2007:Q4 and 2017:Q4, where a missing value will be recorded.

4. The Loss Functions

Following D'Agostino et al., we evaluate point predictions using squared-error loss, that is, the squared forecast error (SFE). As argued by Clements and Hendry (1993a, 1993b), comparisons of forecasts on MSFE (mean SFE) suffer from a number of limitations and shortcomings, and these are inherited by RPS and QPS. Namely, MSFEs lack invariance to nonsingular linear transformations of the forecasts when the forecasts are multivariate (acknowledging that a forecaster at time t typically produces forecasts of both inflation and output growth, either point forecasts or histograms) and/or multiple horizons. Hendry and Martinez (2017) extend the work of Clements and Hendry (1993a) on multivariate measures, and see also Anderson and Vahid (2011) for an application.

It should also be borne in mind that we consider whether a forecaster is better than another, in terms of point forecast performance assessed by SFE, or the quality of their histograms, as assessed by QPS or RPS. Suppose Forecaster A's MSFE is lower than that of Forecaster B. However, it could still be the case that the forecasts of B add value to those of Forecaster A, and this would be the case if A fails to 'forecast encompass' B. Chong and Hendry (1986) develop the concept of forecast encompassing, and Ericsson (1992) discusses the relationship between MSFE dominance (our concern in this paper) and forecast encompassing. The notion of forecast encompassing is typically framed in terms of point forecasting and squared error loss, although Clements and Harvey (2010, 2011) develop tests for probability forecasts, and the concept is applicable to the evaluation of histograms.[10]

For the probability assessments, the loss function we choose is determined in part by the form in which the assessments are reported, namely as histograms. For assessing both point predictions and histograms, the actual values are the advance estimates, so that the rate of GDP growth in 2010, say, is the value available at the end of January 2011, as provided in the Real-time Data Research Center.[11] This seems preferable to using latest-vintage data, and using 'real-time' data is common practice in macro-research. We also check whether our results change if we use the second-quarterly estimates (in our example, the vintage available in 2011:Q2).

The histogram forecasts are assessed using two scoring rules, the quadratic probability score (QPS: Brier 1950) and the ranked probability score (RPS: Epstein 1969). A scoring rule (or loss function) assigns a numerical score based on the density and the realization of the variable. Although the log score is perhaps the most popular scoring rule for densities (see, e.g., Winkler 1967), QPS and RPS have the advantage that they can be calculated directly from the histograms, that is, without making any additional assumptions.[12]

QPS and RPS are defined by:

$$QPS = \sum_{k=1}^{K} \left(p^k - y^k \right)^2 \tag{1}$$

and:

$$RPS = \sum_{k=1}^{K} \left(P^k - Y^k \right)^2 \tag{2}$$

for a single histogram with K bins (indexed by the superscript k), where p^k is the probability assigned to bin k. y^k is an indicator variable equal to 1 when the actual value is in bin k, and zero otherwise. In the definition of RPS, P^k is the cumulative probability (i.e., $P^k = \sum_{s=1}^{k} p^s$), and similarly Y^k cumulates y^s. Note that if $y^{s_1} = 1$, then $Y^k = 1$ for all $k \geq s_1$.[13]

---

[10] We leave an analysis of forecasting encompassing, and multivariate loss measures, for future work.

[11] https://www.philadelphiafed.org/research-and-data/real-time-center.

[12] For example, to calculate the log score we would need to make an assumption about the distribution of the probability mass within each of the histogram intervals, which might be done by first fitting a density to the histogram, see, e.g., (Giordani and Söderlind 2003) and (Engelberg et al. 2009).

[13] The SPF definitions of the bin locations are formally given as, e.g., 4 to 5.9, and 6 to 7.9, and so on. Usually the bins are interpreted as $[4, 6]$ and $[6, 8]$, but if the realization is, say, 5.98, interpreting this as falling in the lower bin ($[4, 6]$) might be

We consider both scores. Being based on cumulative distributions, RPS will penalize less severely forecasts with probability close to the bin containing the actual value, relative to QPS. For QPS, a given probability outside the bin in which the actual falls has the same cost regardless of how near or far it is from the outcome-bin. For this reason, the RPS seems preferable to QPS.

The calculation of QPS and RPS only requires knowledge of the probabilities assigned to each bin, $\left\{p^k\right\}$, provided explicitly by the survey respondents, and a stance on what constitutes the actual value (and therefore y^k). Unlike fitting parametric distributions, no difficulties arise when probability mass is assigned to only one or two bins.

5. Results

We begin by presenting results for point predictions and histogram forecasts for the whole set of forecasters. We then adapt the testing procedure to allow an assessment of whether the findings are affected by forecasts made by 'inexperienced' forecasters. We use first-available actual values throughout to evaluate the forecasts. We end by checking whether the results are qualitatively unchanged if the advance-estimate actuals are replaced with a vintage released two quarters after the reference quarter.

5.1. Results for all Forecasters

Table 2 indicates that the 'best' output growth forecaster occupies that position by chance at all the four horizons we consider. For example, at the shortest horizon ($h = 1$) the loss for the best forecaster ('top') is 0.14, which lies within the 5th and 95th percentiles of the distribution of the top forecaster across the 1000 simulated samples, which are [0.08, 0.40]. Because the bootstrap distribution is generated by random assignment of forecasts to individuals, the actual best value of 0.14 is consistent with chance. The value of 19% indicates that on 19% of the replications a better (lower) value than the actual value was obtained.

For inflation, only at the shortest horizon is there any evidence that the top forecaster is actually more skillful, in that only 3% of the replications yield a smaller loss, whereas at the other horizons the actual is no better than would be expected at conventional significance levels.

For output growth the 25th percentile and median forecasters are generally better than chance would dictate, but little evidence that the 'better forecasters' are in reality any better. The same is broadly true for inflation. D'Agostino et al. (2012) note that better-than-chance performance at lower echelons can be attributed to a small number of very poor forecasters, and experiment with trimming the worst 20% of forecasters. We do not do that here given the smaller number of forecasters resulting from the fixed-event nature of the forecasts, and instead we focus on the upper half of the distribution.

Although not directly comparable to D'Agostino et al. (2012) for the reasons given in Section 2, when they impose a minimum number of forecasts per forecaster (their Table 2), the two sets of results line up for the best forecaster. They do not reject their best GDP forecaster being no better in reality at one quarter or one-year horizons, although their best inflation forecaster is actually better one quarter ahead, but not one year ahead, matching the findings in Table 2.

The results for the histograms in Table 3 provide more evidence that some forecasters really are better than others. The two measures QPS and RPS give broadly equivalent results for the 'better forecasters'. For both inflation and output growth, a more consistent picture emerges than for the point predictions. The results suggest that the better forecasters really are better at the shorter horizons (one and two quarters ahead) for the 'top' forecaster, and the 5th and 25th best forecasters. This is not readily apparent from the number of asterisks in the table: an asterisk * indicates an actual outside the

misleading. In such circumstances we instead assume that $y_t^k = \frac{1}{2}$ for these two bins, and the RPS Y_t^k values for the two bins are $\frac{1}{2}$, 1, respectively.

5th to 95% percentiles of the bootstrap distribution. However, note that the proportion of the bootstrap distribution smaller than the actual is less than 10% for the top forecaster for $h = 1$ for both variables (and both measures), and is less than 5% for the 5th and 25th best at both $h = 1$ and $h = 2$ for both variables. That is, focusing on the comparison of the actual score to a two-sided confidence interval with a fixed level may not tell the whole story. Of interest here is whether the forecasts are better (lower score) than chance would dictate, and this is given by the probability of obtaining a lower score by chance (that is, under the null). Moreover, the use of conventional confidence levels has been criticized in the recent literature—see (Kim and Choi 2017). This is beyond the scope of our paper, but it is worth remarking that an 80% confidence level, for example, would strengthen our findings that forecasters are not the same in terms of the accuracy of their probability assessments.

Table 2. Annual calendar year point predictions (Squared Forecast Error (SFE)): empirical percentiles compared to bootstrap distribution, 1981:Q3–2018:Q4.

h	**Top**			**5**			**25**			**50**		
						Output growth						
4	0.33	[0.14, 0.42]	68%	0.44	[0.35, 0.54]	49%	0.64 *	[0.65, 0.80]	2%	0.85 *	[0.86, 1.01]	4%
3	0.18	[0.08, 0.35]	36%	0.31	[0.30, 0.49]	7%	0.48 *	[0.59, 0.74]	0%	0.76 *	[0.82, 0.98]	0%
2	0.10	[0.09, 0.35]	6%	0.37	[0.33, 0.50]	20%	0.60 *	[0.64, 0.76]	1%	0.77 *	[0.82, 0.96]	0%
1	0.14	[0.08, 0.40]	19%	0.46	[0.37, 0.56]	47%	0.70	[0.66, 0.78]	30%	0.86	[0.84, 0.97]	15%
						Inflation						
4	0.15	[0.06, 0.31]	33%	0.25	[0.25, 0.44]	5%	0.51 *	[0.56, 0.72]	0%	0.70 *	[0.80, 1.01]	0%
3	0.20	[0.08, 0.30]	53%	0.27	[0.24, 0.42]	12%	0.51 *	[0.54, 0.69]	1%	0.80	[0.77, 0.98]	11%
2	0.18	[0.09 , 0.29]	38%	0.27	[0.26, 0.39]	9%	0.49 *	[0.50, 0.65]	3%	0.77	[0.74, 0.94]	15%
1	0.03 *	[0.03, 0.19]	3%	0.19	[0.16, 0.31]	15%	0.40 *	[0.41, 0.55]	4%	0.57 *	[0.64, 0.88]	0%

For each forecast horizon (4, 3, 2, 1) and each of four percentiles ('top', 5, 25, 50), the table shows: (i) the score or loss for the actual forecaster occupying that percentile at that horizon, followed by (ii) the 5th and 95th percentiles for that position calculated from the bootstrap distribution (generated under the assumption of equal accuracy), and (iii) the proportion of the bootstrap distribution less than the actual. An * indicates a score or loss for the actual forecaster that is outside of the 5th to 95th confidence intervals.

Table 3. Annual calendar year histograms: empirical percentiles compared to bootstrap distribution, for RPS and QPS. 1981:Q3–2018:Q4.

h	**Top**			**5**			**25**			50		
						RPS						
						Output growth						
4	0.38	[0.31, 0.60]	17%	0.56	[0.54, 0.70]	12%	0.80	[0.78, 0.88]	19%	0.95	[0.92, 1.02]	28%
3	0.35	[0.18, 0.48]	53%	0.47	[0.44, 0.63]	9%	0.66 *	[0.73, 0.84]	0%	0.90 *	[0.91, 1.02]	2%
2	0.02 *	[0.10, 0.41]	0%	0.36 *	[0.37, 0.56]	3%	0.55 *	[0.67, 0.81]	0%	0.86 *	[0.88, 1.02]	2%
1	0.01	[0.01, 0.24]	7%	0.11 *	[0.17, 0.37]	1%	0.39 *	[0.52, 0.71]	0%	0.75 *	[0.81, 1.02]	0%
						Inflation						
4	0.33	[0.23, 0.52]	26%	0.44 *	[0.45, 0.64]	3%	0.73	[0.73, 0.85]	4%	0.90	[0.90, 1.03]	6%
3	0.31	[0.15, 0.43]	50%	0.38	[0.36, 0.56]	7%	0.63 *	[0.68, 0.81]	0%	0.90	[0.88, 1.01]	18%
2	0.17	[0.05, 0.36]	40%	0.20 *	[0.28, 0.49]	0%	0.53 *	[0.62, 0.77]	0%	0.97	[0.86, 1.02]	78%
1	0.01	[0.01, 0.28]	5%	0.16 *	[0.19, 0.41]	2%	0.46 *	[0.57, 0.74]	0%	0.80 *	[0.82, 1.02]	1%
						QPS						
						Output growth						
4	0.35 *	[0.37, 0.67]	4%	0.61	[0.61, 0.77]	5%	0.87	[0.83, 0.92]	48%	0.99	[0.95, 1.03]	58%
3	0.41	[0.22, 0.55]	55%	0.52	[0.51, 0.70]	6%	0.73 *	[0.78, 0.89]	0%	0.98	[0.94, 1.03]	49%
2	0.02 *	[0.11, 0.46]	0%	0.41 *	[0.42, 0.61]	4%	0.63 *	[0.73, 0.85]	0%	0.96	[0.92, 1.03]	39%
1	0.01	[0.01, 0.27]	5%	0.15 *	[0.19, 0.41]	2%	0.44 *	[0.58, 0.75]	0%	0.83 *	[0.86, 1.04]	2%
						Inflation						
4	0.40	[0.29, 0.60]	23%	0.53 *	[0.53, 0.71]	4%	0.78 *	[0.78, 0.89]	4%	0.93	[0.92, 1.03]	7%
3	0.39	[0.19, 0.51]	58%	0.47	[0.43, 0.63]	15%	0.73 *	[0.75, 0.86]	2%	0.95	[0.92, 1.02]	24%
2	0.18	[0.07, 0.43]	29%	0.28 *	[0.34, 0.57]	1%	0.64 *	[0.69, 0.83]	0%	1.00	[0.90, 1.03]	83%
1	0.02	[0.01, 0.31]	5%	0.19 *	[0.20, 0.46]	3%	0.53 *	[0.62, 0.77]	0%	0.86	[0.86, 1.03]	5%

See the notes for Table 2.

To summarise: the results for the histogram forecasts strongly suggest the better-performing forecasters are actually better. The histogram statistics more frequently detect statistically significant better performance by the 'top' performer, and by the forecasters occupying the 5th and 25th percentiles of the empirical distribution.

5.2. Inexperienced Forecasters

A possible reason for the difference in the findings for the two sets of forecasts is the greater complexity of the task of providing a histogram forecast, which means newcomers may be at a disadvantage compared to experienced survey respondents.[14] Experienced survey respondents may have benefitted from 'learning from doing', and might better understand the task and the form of the required survey return, and have models or methods in place to generate the required histogram forecasts. Of course the same may be true for point predictions, but teething troubles might be more acute for the histogram forecasts especially if these are not ordinarily produced.

To assess the impact of 'inexperience' on our findings, we exclude the forecasts made by forecasters when they were inexperienced. We make this operational as follows. Consider a given survey date, t, which is in quarter Q. The quarter of the year of the survey determines the forecast horizon, given that we are working with fixed-event forecasts (of calendar year growth rates). Consider all those who responded at time t. Of these, the inexperienced forecasts are defined as the forecasts of individuals who responded to none of the previous four-quarter Q surveys (i.e., none of the surveys at $t-4$, $t-8$, $t-12$, and $t-16$). That is, we require the forecaster to have given at least one forecast at the horizon of interest to be experienced. When we score an individual's forecasts, we ignore the inexperienced forecasts. At survey $t+4$, we would include the forecast of an individual whose survey-t forecast was ignored, as this forecast now comes from an individual deemed to be experienced.[15] We have adopted a relatively simple way of defining inexperienced forecasters—other choices could be made and might lead to different results. Our approach provides a simple check on whether inexperience appears to matter.

The approach to classifying forecasts means we lose the first four years of surveys at the beginning of the sample (1981:Q3 to 1985:Q2) for the set of forecasts we can score when we filter our inexperienced forecasts. Hence we check that the results on the shortened sample (1985:Q3 to 2018:Q4), without the inexperienced filter, are qualitatively similar to those on the full sample, and we compare the shortened-sample results directly to those using the filter. This ensures that the results of filtering out the inexperienced are not caused by using a different sample.[16]

The bootstrap test is then applied in exactly the same way as above. When there is a missing value, due to an inexperienced forecast, the corresponding simulated forecaster is also given a missing value. The inexperienced forecasts play no role in terms of calculating the normalization factors for period t, and nor can they be randomly reassigned to other forecasters. Hence the simulated forecasters match the actual forecasters, where the actual no longer contain inexperienced forecasts (on our definition).

Table 1 can again be used to illustrate. The first survey we consider is 1985:Q4—the previous four Q4-surveys are used to determine whether the 1985:Q4 survey forecasts are from experienced forecasters. The forecasts made by both Forecasters 3 and 9 will be included, as these forecasters are deemed experienced. (The score will be normalized using the scores of the other experienced forecasts

---

[14] Clements (2020) analyses whether survey joiners as a group are different from experienced forecasters (and whether leavers are different from those who remain). In this section we briefly touch on this issue when we consider if our findings are sensitive to omitting newcomers.

[15] Although she will have only made one forecast of horizon $5-Q$, she will also likely have made forecasts of other horizons in response to the intervening surveys ($t+1$, $t+2$ and $t+3$), and possibly in response to $t-1$, etc.

[16] When we shorten the sample to exclude the first 16 surveys, or change the effective number of forecasts made by individuals by replacing inexperienced forecasts with missing values, we ensure that there are a minimum number of forecasts made by each included individual (at least 5). This is to reduce the likelihood of some respondents 'getting lucky', discussed in an earlier footnote.

to the 1985:Q4 survey. There are no inexperienced forecasts amongst the 10 forecasters illustrated in the Table). In 1986:Q4, the forecast made by Forecaster 8 is excluded because Forecaster 8 is a newcomer. By the time we get to 1991:Q4, 7 of the 10 individuals register a forecast. Three of the seven are made by newcomers (Forecasters 1, 4 and 5), and so on.

The results are recorded in Table 4 for the point predictions, and Tables 5 and 6 for the histograms scored by RPS and QPS. The results suggest that for both types of forecast filtering out inexperienced forecasts has little effect. Excluding the period 1981:Q3 to 1985:Q2 does not qualitatively change the pattern of results,[17] and comparing the results after filtering out inexperienced forecasts to either the full sample results (Tables 2 and 3) or the shortened sample results, suggests inexperience matters little.

Table 4. Annual calendar year point predictions (SFE): empirical percentiles compared to bootstrap distribution. Using a shortened sample, and excluding relatively inexperienced forecasters.

h	Top			5			25			50		
Output growth: Shortened Sample												
4	0.33	[0.11, 0.41]	71%	0.44	[0.34, 0.54]	52%	0.62 *	[0.64, 0.79]	2%	0.85 *	[0.86, 1.03]	3%
3	0.18	[0.08, 0.32]	38%	0.27	[0.26, 0.45]	7%	0.48 *	[0.58, 0.73]	0%	0.74 *	[0.81, 0.98]	0%
2	0.36	[0.15, 0.42]	74%	0.47	[0.37, 0.54]	59%	0.63 *	[0.65, 0.77]	2%	0.82 *	[0.84, 0.97]	1%
1	0.44	[0.17, 0.48]	87%	0.46	[0.40, 0.58]	24%	0.70	[0.68, 0.80]	16%	0.85	[0.85, 0.98]	8%
Inflation: Shortened Sample												
4	0.07	[0.05, 0.30]	10%	0.24	[0.23, 0.45]	7%	0.50 *	[0.55, 0.72]	1%	0.67 *	[0.80, 1.01]	0%
3	0.22	[0.09, 0.32]	60%	0.28	[0.25, 0.43]	11%	0.49 *	[0.53, 0.70]	1%	0.77 *	[0.78, 0.98]	5%
2	0.18	[0.09, 0.29]	35%	0.27	[0.24, 0.39]	17%	0.49 *	[0.49, 0.65]	4%	0.80	[0.74, 0.97]	26%
1	0.11	[0.05, 0.23]	35%	0.19	[0.18, 0.33]	9%	0.44	[0.42, 0.59]	11%	0.60 *	[0.67, 0.92]	0%
Output growth: Excluding Inexperienced												
4	0.29	[0.14, 0.44]	43%	0.37	[0.32, 0.53]	16%	0.60 *	[0.65, 0.79]	0%	0.86	[0.85, 1.01]	9%
3	0.25	[0.10, 0.36]	58%	0.37	[0.29, 0.49]	39%	0.60	[0.58, 0.75]	9%	0.77 *	[0.81, 1.01]	0%
2	0.28	[0.15, 0.43]	39%	0.42	[0.36, 0.55]	27%	0.62 *	[0.64, 0.77]	2%	0.81 *	[0.83, 0.97]	1%
1	0.37	[0.20, 0.48]	55%	0.52	[0.41, 0.59]	64%	0.68	[0.67, 0.79]	12%	0.82 *	[0.85, 0.98]	1%
Inflation: Excluding Inexperienced												
4	0.08	[0.07, 0.34]	7%	0.18 *	[0.23, 0.45]	1%	0.53 *	[0.57, 0.75]	1%	0.78 *	[0.81, 1.01]	1%
3	0.24	[0.09, 0.35]	61%	0.29	[0.23, 0.44]	24%	0.50 *	[0.54, 0.72]	1%	0.80	[0.78, 1.01]	9%
2	0.10	[0.08, 0.30]	8%	0.26	[0.25, 0.40]	9%	0.47 *	[0.49, 0.65]	1%	0.80	[0.73, 0.96]	26%
1	0.15	[0.05, 0.23]	61%	0.21	[0.18, 0.33]	17%	0.40 *	[0.40, 0.57]	4%	0.58 *	[0.65, 0.92]	0%

See the notes for Table 2.

[17] In terms of inference regarding whether the better forecasters really are better. Note that the loss of the best and better forecasters sometimes changes (e.g., compare the SFE loss for the best $h = 2$ GDP forecaster in Table 2 with that of the best forecaster in Table 4). This might occur when the shortening of the sample changes the identity of the best (or better) forecaster.

Table 5. Annual calendar year histogram forecasts (RPS): empirical percentiles compared to bootstrap distribution. Using a shortened sample, and excluding relatively inexperienced forecasters.

h	Top			5			25			50		
Output growth: Shortened Sample												
4	0.45	[0.32, 0.60]	40%	0.60	[0.54, 0.71]	30%	0.80	[0.78, 0.88]	21%	0.95	[0.92, 1.03]	25%
3	0.32	[0.20, 0.53]	28%	0.47	[0.46, 0.65]	7%	0.67 *	[0.74, 0.86]	0%	0.90 *	[0.91, 1.02]	2%
2	0.34	[0.16, 0.46]	57%	0.36 *	[0.39, 0.58]	1%	0.55 *	[0.68, 0.81]	0%	0.87 *	[0.88, 1.03]	3%
1	0.02	[0.02, 0.29]	5%	0.11 *	[0.22, 0.42]	0%	0.41 *	[0.56, 0.75]	0%	0.76 *	[0.82, 1.03]	0%
Inflation: Shortened Sample												
4	0.33	[0.22, 0.52]	25%	0.47	[0.45, 0.64]	7%	0.72	[0.72, 0.86]	5%	0.90 *	[0.91, 1.04]	4%
3	0.28	[0.16, 0.48]	33%	0.36 *	[0.39, 0.59]	2%	0.67 *	[0.70, 0.83]	1%	0.91	[0.88, 1.03]	13%
2	0.14	[0.05, 0.38]	27%	0.27 *	[0.32, 0.53]	1%	0.53 *	[0.64, 0.79]	0%	0.96	[0.86, 1.03]	66%
1	0.06	[0.03, 0.31]	11%	0.17 *	[0.24, 0.46]	0%	0.46 *	[0.57, 0.75]	0%	0.80 *	[0.82, 1.02]	1%
Output growth:Excluding Inexperienced												
4	0.45	[0.32, 0.62]	34%	0.53	[0.51, 0.71]	8%	0.81	[0.78, 0.89]	23%	0.95	[0.92, 1.03]	22%
3	0.31	[0.17, 0.55]	28%	0.45	[0.41, 0.65]	10%	0.66 *	[0.75, 0.87]	0%	0.91 *	[0.91, 1.03]	4%
2	0.27	[0.19, 0.48]	20%	0.36 *	[0.41, 0.60]	1%	0.53 *	[0.68, 0.82]	0%	0.83 *	[0.88, 1.03]	0%
1	0.02	[0.02, 0.30]	6%	0.10 *	[0.15, 0.41]	1%	0.32 *	[0.54, 0.74]	0%	0.76 *	[0.81, 1.03]	1%
Inflation: Excluding Inexperienced												
4	0.19 *	[0.24, 0.56]	2%	0.41 *	[0.43, 0.65]	2%	0.69 *	[0.73, 0.87]	0%	0.91	[0.90, 1.04]	7%
3	0.25	[0.17, 0.51]	18%	0.36 *	[0.37, 0.60]	3%	0.60 *	[0.68, 0.83]	0%	0.92	[0.88, 1.04]	19%
2	0.13	[0.05, 0.40]	21%	0.20 *	[0.24, 0.50]	1%	0.47 *	[0.62, 0.79]	0%	0.90	[0.85, 1.04]	19%
1	0.06	[0.02, 0.32]	19%	0.16 *	[0.18, 0.44]	3%	0.47 *	[0.57, 0.76]	0%	0.77 *	[0.82, 1.03]	0%

See the notes for Table 2.

Table 6. Annual Calendar Year Histogram Forecasts (QPS): empirical percentiles compared to bootstrap distribution. Using a shortened sample, and excluding relatively inexperienced forecasters.

h	Top			5			25			50		
Output growth: Shortened Sample												
4	0.54	[0.38, 0.68]	42%	0.62 *	[0.62, 0.77]	4%	0.85	[0.83, 0.92]	18%	0.98	[0.95, 1.03]	33%
3	0.33	[0.24, 0.59]	17%	0.58	[0.53, 0.70]	22%	0.78 *	[0.79, 0.89]	2%	0.98	[0.94, 1.03]	50%
2	0.36	[0.17, 0.52]	45%	0.42 *	[0.44, 0.64]	2%	0.62 *	[0.73, 0.86]	0%	0.96	[0.91, 1.04]	31%
1	0.02 *	[0.02, 0.32]	4%	0.11 *	[0.24, 0.47]	0%	0.45 *	[0.60, 0.78]	0%	0.84 *	[0.85, 1.04]	3%
Inflation: Shortened Sample												
4	0.40	[0.29, 0.60]	22%	0.55	[0.55, 0.72]	7%	0.78 *	[0.78, 0.89]	4%	0.94	[0.93, 1.04]	8%
3	0.31	[0.19, 0.54]	28%	0.46	[0.45, 0.67]	6%	0.78	[0.76, 0.88]	9%	0.95	[0.92, 1.04]	22%
2	0.13	[0.06, 0.46]	18%	0.30 *	[0.38, 0.61]	0%	0.64 *	[0.71, 0.84]	0%	1.01	[0.90, 1.04]	82%
1	0.07	[0.04, 0.36]	13%	0.19 *	[0.28, 0.52]	0%	0.53 *	[0.62, 0.79]	0%	0.86	[0.86, 1.04]	5%
Output growth:Excluding Inexperienced												
4	0.51	[0.38, 0.70]	30%	0.62	[0.59, 0.77]	13%	0.88	[0.83, 0.92]	46%	1.01	[0.95, 1.03]	73%
3	0.32	[0.20, 0.61]	19%	0.54	[0.48, 0.71]	20%	0.77 *	[0.80, 0.90]	1%	0.95	[0.94, 1.04]	13%
2	0.29	[0.22, 0.54]	14%	0.42 *	[0.47, 0.66]	1%	0.58 *	[0.74, 0.87]	0%	0.92	[0.92, 1.05]	6%
1	0.02	[0.02, 0.33]	5%	0.12 *	[0.17, 0.46]	1%	0.34 *	[0.59, 0.78]	0%	0.83 *	[0.86, 1.05]	2%
Inflation: Excluding Inexperienced												
4	0.31	[0.31, 0.63]	5%	0.49 *	[0.52, 0.72]	2%	0.77 *	[0.79, 0.90]	2%	0.92 *	[0.93, 1.03]	2%
3	0.28	[0.19, 0.58]	18%	0.46	[0.42, 0.67]	10%	0.70 *	[0.75, 0.88]	0%	0.97	[0.92, 1.05]	42%
2	0.12	[0.07, 0.48]	13%	0.21 *	[0.30, 0.59]	1%	0.56 *	[0.69, 0.84]	0%	0.99	[0.89, 1.04]	73%
1	0.07	[0.02, 0.37]	21%	0.17 *	[0.20, 0.49]	3%	0.52 *	[0.62, 0.80]	0%	0.83 *	[0.86, 1.04]	1%

See the notes for Table 2.

5.3. Sensitivity of Results to Choice of Actual Values

As a check that the results are not sensitive to the precise choice of real-time actual values (the results reported so far are based on 'advance' estimates), we re-ran the bootstrap test for the point predictions and for the histograms on the whole sample, using the second-quarterly release values as actual values. The results are reported in Table 7. A comparison of the results for point predictions to

those in Table 2, and of the results for the histograms to Table 3, suggests that changing the vintage of the actuals does not qualitatively affect our findings.

Table 7. Annual calendar year point predictions and histograms: empirical percentiles compared to bootstrap distribution, 1981:Q3—2018:Q4. Using actual available two-quarters after the reference year.

h	Top			5			25			50		
Output growth: Point Predictions (SFE)												
4	0.32	[0.15, 0.42]	66%	0.44	[0.35, 0.53]	45%	0.63 *	[0.65, 0.80]	1%	0.84 *	[0.86, 1.01]	3%
3	0.24	[0.08, 0.34]	63%	0.33	[0.30, 0.49]	10%	0.52 *	[0.59, 0.74]	0%	0.80 *	[0.81, 0.99]	2%
2	0.12	[0.11, 0.36]	6%	0.40	[0.34, 0.51]	31%	0.62 *	[0.64, 0.76]	1%	0.80 *	[0.82, 0.96]	1%
1	0.27	[0.09, 0.42]	57%	0.50	[0.39, 0.57]	60%	0.71	[0.67, 0.78]	30%	0.90	[0.84, 0.97]	47%
Inflation: Point Predictions (SFE)												
4	0.16	[0.06, 0.31]	38%	0.25	[0.24, 0.44]	7%	0.50 *	[0.56, 0.72]	1%	0.70 *	[0.80, 1.01]	0%
3	0.20	[0.08, 0.30]	57%	0.24 *	[0.24, 0.42]	4%	0.50 *	[0.54, 0.69]	1%	0.80	[0.77, 0.98]	15%
2	0.22	[0.10, 0.30]	59%	0.32	[0.27, 0.40]	33%	0.50 *	[0.51, 0.65]	3%	0.79	[0.74, 0.95]	21%
1	0.05	[0.03, 0.17]	13%	0.18	[0.15, 0.29]	23%	0.41	[0.40, 0.55]	7%	0.57 *	[0.65, 0.87]	0%
Output growth: Histograms (RPS)												
4	0.38	[0.31, 0.60]	17%	0.54	[0.54, 0.70]	5%	0.80	[0.78, 0.88]	22%	0.94	[0.92, 1.02]	20%
3	0.34	[0.19, 0.48]	47%	0.45 *	[0.45, 0.63]	4%	0.67 *	[0.74, 0.85]	0%	0.92	[0.91, 1.02]	8%
2	0.02 *	[0.10, 0.41]	0%	0.38 *	[0.39, 0.57]	4%	0.57 *	[0.67, 0.81]	0%	0.88	[0.88, 1.03]	5%
1	0.01	[0.01, 0.25]	7%	0.11 *	[0.18, 0.39]	1%	0.45 *	[0.54, 0.72]	0%	0.82	[0.82, 1.02]	5%
Inflation: Histograms (RPS)												
4	0.35	[0.24, 0.52]	31%	0.44 *	[0.46, 0.65]	3%	0.72 *	[0.73, 0.85]	4%	0.89 *	[0.90, 1.03]	3%
3	0.31	[0.17, 0.46]	45%	0.40	[0.39, 0.57]	7%	0.67 *	[0.68, 0.81]	1%	0.90	[0.87, 1.01]	17%
2	0.18	[0.08, 0.37]	30%	0.23 *	[0.29, 0.50]	1%	0.55 *	[0.63, 0.77]	0%	0.96	[0.86, 1.02]	71%
1	0.01 *	[0.02, 0.28]	1%	0.15 *	[0.20, 0.41]	1%	0.42 *	[0.56, 0.73]	0%	0.85	[0.82, 1.02]	11%
Output growth: Histograms (QPS)												
4	0.35 *	[0.36, 0.68]	4%	0.63	[0.61, 0.77]	12%	0.87	[0.83, 0.91]	50%	1.00	[0.95, 1.02]	75%
3	0.39	[0.22, 0.55]	45%	0.52	[0.52, 0.70]	5%	0.74 *	[0.79, 0.89]	0%	0.98	[0.94, 1.03]	45%
2	0.02 *	[0.11, 0.46]	0%	0.44	[0.43, 0.62]	7%	0.63 *	[0.73, 0.85]	0%	0.98	[0.91, 1.03]	57%
1	0.01 *	[0.01, 0.28]	4%	0.15 *	[0.20, 0.43]	1%	0.48 *	[0.58, 0.75]	0%	0.86	[0.86, 1.04]	7%
Inflation: Histograms (QPS)												
4	0.43	[0.29, 0.60]	30%	0.50 *	[0.54, 0.72]	1%	0.78 *	[0.78, 0.89]	3%	0.93	[0.93, 1.03]	8%
3	0.44	[0.22, 0.52]	71%	0.49	[0.45, 0.64]	13%	0.75	[0.75, 0.87]	6%	0.97	[0.92, 1.02]	45%
2	0.18	[0.10, 0.43]	20%	0.30 *	[0.36, 0.57]	1%	0.62 *	[0.70, 0.83]	0%	0.99	[0.90, 1.03]	74%
1	0.01 *	[0.02, 0.31]	0%	0.21 *	[0.22, 0.46]	3%	0.53 *	[0.61, 0.77]	0%	0.88	[0.86, 1.03]	12%

See the notes for Table 2.

6. Related Literature

To the best of my knowledge there are no studies attempting to determine whether some survey forecasters' probability assessments are really better than those of others. The results in this paper are the first to bear on this question.

In terms of point predictions, Clements (2019a) considers the set of forecasters who made the most forecasts over a given period. The ranks of the forecasters are calculated over two sub-periods (using normalized forecast errors), and the null of equal accuracy is interpreted to mean a zero correlation between the ranks in the two periods. That is, that there is no persistence in the relative performance of the forecasters over the two sub-periods. The approach is based on the simple idea that if there are real differences between forecasters, then the ranking of forecasters in one period ought to be informative about the ranking in a subsequent period. Clements (2019a) considers different variables, consumers' expenditure growth, investment growth and real GDP growth jointly, using either the trace or determinant of the second moment matrix, of either one or four-step ahead quarterly forecast errors. The Spearman test of no correlation in the ranks is rejected at conventional significant levels. As well as the different set of variables under consideration, the study uses the quarterly series of

fixed-horizon forecasts, and so uses a relatively long span of forecast data compared to the annual series we have considered.

There would appear to be advantages and disadvantages to both approaches: the bootstrap test, and the rank correlation test across sub-samples. The bootstrap test considers the forecasts en masse,[18] whereas the requirement that an individual makes a reasonable number of forecasts in each of the two sub-periods leads to a focus on a smaller number of forecasters: Clements (2019a) considers the 50 most prolific respondents. The most prolific might not be typical of the wider population of forecasters, so the null is not the same as in the bootstrap test. The test is of whether within in a given sample of forecasters (the most prolific) some appear to perform consistently better than others. The results of the bootstrap test depend on all forecasters (subject to the caveat above about exclusion) and are liable to being unduly influenced by a small number of very poor forecasts.

In terms of implementation, the bootstrap test simply requires choosing the minimum number of forecasts required for an individual to be included. The sub-sample approach requires making a decision on the period for splitting the sample into two, and the minimum number of forecasts in each of the two samples required for an individual to be included. Table 8 illustrates the sub-sample approach, where we split the sample in the middle, into 1981:3 to 1999:4, and 2000:1 to 2018:4, and require that each respondent makes a minimum of 15 forecasts in each sub-period. This gives 17/18 respondents for the tests of the histograms, and 19/20 for the point predictions. We bundle the forecasts made in the four quarters of the year together, that is, the S_i average over forecasts of horizons of one quarter to one year ahead.[19] The test is based on a comparison of the ranks of S_i^1 and S_i^2, where the superscript denotes the sub-sample, and i indexes the eligible individuals.

Table 8. Rank correlation tests of equal accuracy.

	Histograms: RPS		**Point Predictions**	
	GDP	**Inflation**	**GDP**	**Inflation**
r	0.76	0.58	0.37	0.68
p-value	0.00	0.02	0.10	0.00

We test whether the ranking of forecasters during the period 1981:3 to 1999:4 is the same as during the later period 2000:1 to 2018:4. The Spearman rank correlation r lies between -1 and 1, and 0 indicates no relationship. For each test, there are two entries. The first row entry is the small-sample test statistic $r = 1 - \frac{6R}{N\left(N^2-1\right)}$, where R is the sum of squared differences between the ranks (e.g., of the forecasters in the first sample, and in the second sample). At the 5% level, the two-sided critical values for the small-sample statistic are ± 0.485. (This is correct when $N = 17$, as is the case for the inflation histogram forecasts. For the point predictions there are a couple more individuals, and in that case the critical value is conservative). For large samples, the test statistic $\frac{6R - N\left(N^2-1\right)}{N(N+1)\sqrt{N-1}}$ is standard normal. We square the test statistic, and report the probability of obtaining a larger value relative to a chi-squared (one degree of freedom) reference distribution. This corresponds to a two-sided test p-value.

The top panel of Table 8 shows that we reject the null of uncorrelated ranks at conventional significance levels for the histogram forecasts (scored by RPS), and for the inflation point forecasts, but not for the GDP growth point predictions.

Broadly, the two approaches are in tune in suggesting that there is more evidence that there are real differences between individuals in terms of their ability to make accurate probability assessments. However, for the reasons we have explained, it need not be the case that the two tests agree, as the tests emphasize different facets of the inter-forecaster comparisons (e.g., the behaviour of the most prolific forecasters, or of the top forecaster, or the median forecaster, and so on).

---

[18] Possibly after requiring the forecasters to have made a minimum number of responses, and/or trimming the set by ignoring the worst performers.

[19] If we considered each horizon separately, an individual forecaster would need to respond to surveys across 30 years to satisfy our requirements. There would be too few forecasters to obtain meaningful results.

7. Conclusions

D'Agostino et al. (2012) propose a bootstrap test to ascertain whether some survey respondents' forecasts really are more accurate than those of others. They suggest that the actual forecasters in the upper half of the empirical distribution occupy those positions by chance. Using annual series of calendar-year growth rates of real GDP and the GDP deflator, we broadly confirm their findings, which were based on quarterly series of quarterly growth rates. We then present novel empirical evidence regarding whether some forecasters are able to make superior probability assessments to others, and find more evidence that this is the case (compared to when they make point predictions).

Both the probability assessments (in the form of histograms) and the point predictions are of annual calendar year growth rates, at horizons of 1 to 4 quarters ahead. Hence the forecast target and horizons match, and so the different conclusions we draw—regarding whether some forecasters really are better than others—are attributable to the differences in the type of forecast, and not simply because of different forecast horizons, etc. Given that providing a histogram forecast is likely to be more costly, in terms of knowledge acquisition, and the processing of information, it is perhaps not surprising that there are real differences between individuals, reflecting different amounts of resources being devoted to the task.

We investigate whether inexperience affects the results. If we exclude histogram forecasts made by respondents when they were newcomers, and apply the bootstrap test to forecasts made when they were more experienced, are the results unchanged? We find that 'real' differences in histogram forecaster accuracy are still apparent.

Findings based on comparing the rankings of forecasters across two sub-samples tell a similar story: there are persistent differences between individuals in terms of their ability to make accurate histogram forecasts.

Funding: This research received no external funding.

Acknowledgments: Helpful comments from a reviewer of this journal are gratefully acknowledged, as are those of the Special Issue editor, Neil Ericsson. It is an honour to contribute to an issue celebrating the scholarly achievements of David Hendry. I have been truly fortunate to have had David as my DPhil supervisor, and to have had the opportunity to work with him subsequently.

Conflicts of Interest: The authors declare no conflict of interest.

References

Anderson, Heather M., and Farshid Vahid. 2011. VARs, cointegration and common cycle restrictions, chapter 1. In *The Oxford Handbook of Economic Forecasting*. Edited by Michael P. Clements and David F. Hendry. Oxford: Oxford University Press, pp. 9–34.

Brier, Glenn W. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 75: 1–3. [CrossRef]

Castle, Jennifer L., Michael P. Clements, and David F. Hendry. 2019. *Forecasting: An Essential Introduction*. London: Yale University Press.

Chatfield, Chris. 1993. Calculating interval forecasts. *Journal of Business & Economic Statistics* 11: 121–35.

Chong, Yock Y., and David F. Hendry. 1986. Econometric evaluation of linear macro-economic models. *Review of Economic Studies* 53: 671–90. [CrossRef]

Clark, Todd E. 2011. Real-time density forecasts from Bayesian Vector Autoregressions with Stochastic Volatility. *Journal of Business and Economic Statistics* 29: 327–41. [CrossRef]

Clements, Michael P. 2009. Internal consistency of survey respondents' forecasts: Evidence based on the Survey of Professional Forecasters. In *The Methodology and Practice of Econometrics. A Festschrift in Honour of David F. Hendry*. Edited by Jennifer L. Castle and Neil Shephard. Oxford: Oxford University Press, Chapter 8, pp. 206–26.

Clements, Michael P. 2010. Explanations of the Inconsistencies in Survey Respondents Forecasts. *European Economic Review* 54: 536–49. [CrossRef]

Clements, Michael P. 2014a. Forecast Uncertainty—Ex Ante and Ex Post: US Inflation and Output Growth. *Journal of Business & Economic Statistics* 32: 206–16. [CrossRef]

Clements, Michael P. 2014b. Probability distributions or point predictions? Survey forecasts of US output growth and inflation. *International Journal of Forecasting* 30: 99–117. [CrossRef]

Clements, Michael P. 2018. Are macroeconomic density forecasts informative? *International Journal of Forecasting* 34: 181–98. [CrossRef]

Clements, Michael P. 2019a. *Forecaster Efficiency, Accuracy and Disagreement: Evidence Using Individual-Level Survey Data*. Mimeo: ICMA Centre, University of Reading.

Clements, Michael P. 2019b. *Macroeconomic Survey Expectations*. Palgrave Texts in Econometrics. London: Palgrave Macmillan.

Clements, Michael P. 2020. *Do Survey Joiners and Leavers Differ from Regular Participants? The US SPF GDP Growth and Inflation Forecasts*. Mimeo: ICMA Centre, University of Reading.

Clements, Michael P., and David F. Hendry. 1993a. On the limitations of comparing mean squared forecast errors. *Journal of Forecasting* 12: 617–37. [CrossRef]

Clements, Michael P., and David F. Hendry. 1993b. On the limitations of comparing mean squared forecast errors: A reply. *Journal of Forecasting* 12: 669–76. [CrossRef]

Clements, Michael P., and David I. Harvey. 2010. Forecast encompassing tests and probability forecasts. *Journal of Applied Econometrics* 25: 1028–62. [CrossRef]

Clements, Michael P., and David I. Harvey. 2011. Combining probability forecasts. *International Journal of Forecasting* 27: 208–23. [CrossRef]

Coibion, Olivier, and Yuriy Gorodnichenko. 2012. What can survey forecasts tell us about information rigidities? *Journal of Political Economy* 120: 116–59. [CrossRef]

Coibion, Olivier, and Yuriy Gorodnichenko. 2015. Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts. *American Economic Review* 105: 2644–78. [CrossRef]

Croushore, Dean. 1993. Introducing: The Survey of Professional Forecasters. *Federal Reserve Bank of Philadelphia Business Review* 6: 3–15.

D'Agostino, Antonello, Kieran McQuinn, and Karl Whelan. 2012. Are some forecasters really better than others? *Journal of Money, Credit and Banking* 44: 715–32. [CrossRef]

D'Amico, Stefania, and Athanasios Orphanides. 2008. *Uncertainty and Disagreement in Economic Forecasting*. Finance and Economics Discussion Series 2008-56. Washington, DC: Board of Governors of the Federal Reserve System.

D'Amico, Stefania, and Athanasios Orphanides. 2014. *Inflation Uncertainty and Disagreement in Bond Risk Premia*. Working Paper Series WP-2014-24. Chicago: Federal Reserve Bank of Chicago.

Engelberg, Joseph, Charles F. Manski, and Jared Williams. 2011. Assessing the temporal variation of macroeconomic forecasts by a panel of changing composition. *Journal of Applied Econometrics* 26: 1059–78. [CrossRef]

Engelberg, Joseph, Charles F. Manski, and Jared Williams. 2009. Comparing the point predictions and subjective probability distributions of professional forecasters. *Journal of Business and Economic Statistics* 27: 30–41. [CrossRef]

Epstein, Edward S. 1969. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology* 8: 985–87. [CrossRef]

Ericsson, Neil R. 1992. Parameter constancy, mean square forecast errors, and measuring forecast performance: An exposition, extensions, and illustration. *Journal of Policy Modeling* 14: 465–95. [CrossRef]

Giordani, Paolo, and Paul Söderlind. 2003. Inflation forecast uncertainty. *European Economic Review* 47: 1037–59. [CrossRef]

Granger, Clive W. J., and Paul Newbold. 1986. *Forecasting Economic Time Series*, 2nd ed. New York: Academic Press.

Hendry, David F., and Andrew B. Martinez. 2017. Evaluating Multi-Step System Forecasts with Relatively Few Forecast-Error Observations. *International Journal of Forecasting* 33: 359–72. [CrossRef]

Kim, Jae H., and In Choi. 2017. Unit Roots in Economic and Financial Time Series: A Re-Evaluation at the Decision-Based Significance Levels. *Econometrics* 5: 41. [CrossRef]

Knüppel, Malte, and Andreea L. Vladu. 2016. *Approximating Fixed-Horizon Forecasts Using Fixed-Event Forecasts*. Discussion Papers 28/2016. Frankfurt: Deutsche Bundesbank, Research Centre.

McConnell, Margaret M., and Gabriel Perez-Quiros. 2000. Output fluctuations in the United States: What has changed since the early 1980s? *American Economic Review* 90: 1464–76. [CrossRef]

Sims, Christopher A. 2003. Implications of rational inattention. *Journal of Monetary Economics* 50: 665–90. [CrossRef]

Tödter, Karl-Heinz. 2010. *How Useful Is the Carry-Over Effect for Short-Term Economic Forecasting?* Discussion Paper Series 1: Economic Studies, 21/2010. Frankfurt: Deutsche Bundesbank, Research Centre.

Winkler, Robert L. 1967. The quantification of judgement: Some methodological suggestions. *Journal of the American Statistical Associaton* 62: 1105–20. [CrossRef]

Woodford, Michael. 2002. Imperfect common knowledge and the effects of monetary policy. In *Knowledge, Information, and Expectations in Modern Macroeconomics: In Honor of Edmund Phelps*. Edited by Philippe Aghion, Roman Frydman, Joseph Stiglitz and Michael Woodford. Princeton: Princeton University Press, pp. 25–58.