Received: 28 January 2022 | Accepted: 2 February 2022

DOI: 10.1002/ffo2.124

ORIGINAL ARTICLE

WILEY

Point, interval, and density forecasts: Differences in bias, judgment noise, and overall accuracy

Xiaoxiao Niu | Nigel Harvey

Department of Experimental Psychology, University College London, London, UK

Correspondence
Nigel Harvey, Department of Experimental Psychology, University College London, Gower St, London WC1E 6BT, UK.
Email: n.harvey@ucl.ac.uk

Funding information
None

Abstract

There are three main ways in which judgmental predictions are expressed: point forecasts; interval forecasts; probability density forecasts. Do these approaches differ solely in terms of their simplicity of elicitation and the detail they provide? We examined error in values of the central tendency extracted from these three types of forecast in a domain in which all of them are used: lay forecasts of inflation. A first experiment using a between-participant design showed that the mean level of forecasts and the bias in them are unaffected by the type of forecast but that judgment noise (and, hence, overall error) is higher in point forecasts than in interval or density forecasts. A second experiment replicated the difference between point and interval forecasts in a within-participant design (of the sort used in inflation surveys) and showed no effect of the order in which different types of forecast are made but revealed that people are more overconfident in interval than in point forecasts. A third experiment showed that volatility in past data increases bias in point but not interval forecasts, and that taking the average of two point forecasts made by an individual reduces judgment noise to the level found in interval forecasting.

KEYWORDS
forecast error, judgment bias, judgment noise, overconfidence, uncertainty

1 | INTRODUCTION

There are three main ways in which people use judgment to make predictions about the future values of a variable. In point forecasting, they make a single point estimate of its expected value. In interval (or range) forecasting, they provide a range of values within which they judge there is some probability (e.g., 90%) of the outcome occurring. The mid-point of the bounds of the interval is taken to correspond to their expected value of the variable; it should be the same as the point forecast.[1] In probability density forecasting, they provide a probability that the outcome will be in each of a number of different ranges. The mean or median of the distribution of these probabilities should correspond to the expected value of the outcome (i.e., equal to the point forecast).

These types of judgmental forecast vary in two ways: they differ in how simple they are to elicit and in terms of how much information they provide to users. Unlike point forecasts, interval and density forecasts provide users with information about forecasters' uncertainty in their forecasts: in many applications, this is important for planning purposes. Furthermore, density forecasts provide more detail about this uncertainty than interval forecasts. In many domains, this additional information is useful for guiding future decisions.

For example, those assessing financial risks may need to know more than that there is a difference in the width of the interval forecasts for the future returns on two investments; they may wish to know whether it reflects a difference in variance or kurtosis of the distribution of those returns.

There is a trade-off between the simplicity of elicitation of forecasts and the detail that those forecasts provide. In some domains, time pressure arising from the number of forecasts required within a short period of time limits forecasters to the provision of point forecasts: for example, demand forecasters may need to make forecasts for many stock-keeping units within a short period. In other areas, such as meteorology, agriculture, and the nuclear industry, users need density forecasts but time pressure is less of a consideration. Reviewers have tended to focus on either point and interval forecasting (e.g., Lawrence et al., 2006) or on density forecasting (e.g., O'Hagan et al., 2006), partly because the salient issues for research and practice depend on the type of forecast under consideration. Perhaps, for this reason, there has been little concern with comparing the accuracy of the different types of forecast. It is this neglected issue that we focus on here.

Estimates of expected value can be extracted from all three types of forecast: we should expect point forecasts, the mid-point of the bounds describing interval forecasts, and the mean or median of density forecasts to be the same when they are all based on the same data. Here we ask people to make a number of forecasts from the same set of data series. For each person, we measure the mean central (expected) values that they produce for each type of forecast (point, interval, density) to determine whether there is any difference between them. As we also have the true outcomes corresponding to each forecast, we also examine whether the overall accuracy, measured by the root mean-squared error (RMSE), varies across the different types of forecast. This overall measure of error can be decomposed into mean error (ME), also known as bias, directional error, or constant error, and variable error (VE), also known as noise, or inconsistency. Hence, we also examine whether these components of overall error vary across different types of forecast.

Our investigation is framed in terms of inflation forecasting. There were three reasons for this. First, this is one of the few domains in which all three types of forecast (point, interval, and density) are used in practice: surveys of both experts (economists and professional forecasters) and lay people (consumers and households) require them. Second, and related to the first point, inflation forecasting by lay people provides, to the best of our knowledge, the only previous studies that compare different types of forecasting (Bruine de Bruin, Manski, et al., 2011). Third, data series for inflation are available and are regularly updated. Thus, it is possible to provide people with inflation data series for various countries, ask them to make inflation forecasts for those countries, and then compare their forecasts with true outcomes when those are available.

Though our studies are framed within the inflation forecasting domain,[2] we anticipate that our conclusions about differences between different types of forecasting will generalize across all content areas. Indeed, calibration of inflation forecasts and of calibration of other types of forecast (e.g., Benson & Önkal, 1992) are affected in a similar way by specific factors such as feedback (Niu & Harvey, 2022).

2 | RATIONALE AND HYPOTHESES

Bruine de Bruin, Manski, et al. (2011) found that, on average, point forecasts are the same as the mean (and median) value of density forecasts. They obtained this result using a within-participant design: people first made an interval or point forecast, then, if they had made an interval forecast, they made a point forecast, and, finally, they made a density forecast. Here, we seek to replicate this finding using a between-participants design: separate groups of participants made point forecasts, interval forecasts, and density forecasts.

The reason that we made this change is that context effects are known to influence responses in both traditional and online surveys (e.g., Reips, 2002; Smyth et al., 2009; Tourangeau et al., 2000), including surveys of inflation expectations (Niu & Harvey, 2021). Our concern here is that people's responses to survey questions eliciting density forecasts may be influenced by their earlier responses to survey questions eliciting point (or interval) forecasts. In particular, point forecasts may act as mental anchors for estimates of the means of density forecasts: because of under-adjustment (Tversky & Kahneman, 1974), these two values would then be more similar than they would otherwise be. Once the possibility of anchoring is eliminated by the use of a between-participants design, differences between point forecasts, the mid-point of interval forecasts, and the mean value of density forecasts may appear.

Although Bruine de Bruin, Manski, et al. (2011) did not report whether the mid-point of the range (i.e., interval) forecast matched the point forecast and the mean of the density forecast, it is reasonable to assume that it would do so given that they found that the latter two values were the same. Thus, the first hypothesis that we test is the following one:

H_1: Point forecast = Mid-point of the interval forecast = Mean of the density forecast.

This is a null hypothesis. We use a well-powered experiment to examine whether we can obtain evidence inconsistent with it in a between-participants design that excludes the possibility of anchoring effects.

To test our other hypotheses, we extracted from the data the three error measures that we mentioned above. Given that D is the judged rate of inflation for a particular country minus the actual rate of inflation for that country and given that each participant makes inflation judgments for n countries, each participant's ME is given by $\Sigma D/n$. Their VE is given by $\sqrt{(\Sigma(D - \mathrm{ME})^2/n)}$. Their RMSE is given by $\sqrt{(\Sigma(D^2)/n)}$. Equivalently, RMSE can be expressed via its decomposition into ME and VE as $\sqrt{((\mathrm{ME})^2 + (\mathrm{VE})^2)}$.[3]

Many previous studies (e.g., Bruine de Bruin, van der Klaauwm, et al., 2011; Bryan & Venkatu, 2001a, 2001b; Georganas et al., 2014)

have shown that people tend to overforecast inflation. It appears that people have a general expectation that inflation will be higher than it turns out to be. Thus, we test whether the ME is positive.

$\mathbf{H_2}$: ME > 0.

If H_1 is true, it also means that the three different types of forecast will be equally biased. Thus, we also seek to obtain evidence against the following hypothesis:

$\mathbf{H_3}$: $ME_{Point} = ME_{Interval} = ME_{Density}$.

People's judgments are noisy (Kahneman et al., 2021): they are subject not only to bias but to moment-to-moment random variation. It is not surprising, therefore, that taking the average of a number of judgments from a single person produces a more accurate estimate than using a single judgment (Herzog & Hertwig, 2009, 2014; Vul & Pashler, 2008). When someone makes a point forecast, they make a single judgment (f). When they make an interval forecast, they make two judgments ($f + \delta f$; $f - \delta f$). Thus, they estimate f twice; we, therefore, expect the average of the bounds of the interval used to express the range forecast to be more accurate than the point forecast. An analogous argument leads us to expect the central tendency of a density forecast to be more accurate than the mean of the bounds of a range forecast. To examine the validity of these arguments, we test the following hypothesis:

$\mathbf{H_4}$: $VE_{Point} > VE_{Interval} > VE_{Density}$.

If H_3 and H_4 are true, then we should also expect:

$\mathbf{H_5}$: $RMSE_{Point} > RMSE_{Interval} > RMSE_{Density}$.

2.1 | Experiment 1

2.1.1 | Method

Separate groups of participants made point forecasts, interval forecasts, and density forecasts.

Participants

One hundred and thirty-nine participants (75 males, 64 females) with a mean age of 22 years (SD = 5 years) were recruited for the online study. They were divided into three groups: a point forecasting group ($N = 56$); an interval forecasting group ($N = 42$); a density forecasting group ($N = 41$). Forty of these participants were recruited from the participant pool at University College London (UCL) and given 0.25 credits for their participation. The remaining 99 participants were recruited in UCL or China; the former received £3.00 and the latter received 3RMB for taking part. Data were collected between July 1, 2019 and September 30, 2019.

Stimulus materials

Participants in each group were shown 10 graphs of real inflation rate data from 10 different countries. Each one displayed a time series representing 20 years of annual historical inflation data from 1998 to 2017.[4] The last displayed data point was for the period immediately before the one to be forecast. The identities of the 10 countries were not specified; instead, they were labeled with numbers. Seven of the series showed no trends; three contained shallow trends.

The way that data are graphed can affect the forecasts that people make (Lawrence & O'Connor, 1992). For example, people are less likely to follow an upward trend in the data when the last data point is already close to the top of the vertical axis. To avoid such problems, the inflation data series were displayed in the central part of the y-axis scale, which ranged from 12% to −8% and the final points for all 10 countries were in the middle of that scale, ranging from −0.31% to 7.55%. Series were broadly comparable with a mean inflation level at 2.61% (SD = 2.90%) across the 10 series. A typical series is shown in Figure 1.

Two versions of the experiment were programmed, one in English for English speakers and one in Chinese Mandarin for Chinese speakers. To ensure these were comparable, the English version was initially translated into Chinese and then back-translated into English. The back-translation was then compared to the original version to ensure that they matched.

Design

The experiment used a between-participants design. Participants from each language group were randomly assigned to one of three groups: point forecasting, interval forecasting, density forecasting.

Procedure

Participants first saw an information screen that outlined the nature of the study and a consent screen that detailed the ethical permission that had been provided and that elicited their consent for participating. They were then asked basic demographical questions that required them to specify their age, gender, level of education, main academic discipline that they had studied, the country that they had lived in for most of their life, and any economics-related work experience. A brief explanation of the nature of inflation was then provided. After that, participants in each group completed their 10 forecasts. The 10 countries for which inflation had to be forecast were presented in a different random order for each participant.[5]

Instructions given to those in the point forecasting group were: "Below is a series of inflation rates for one country. WHAT WILL HAPPEN NEXT? Please estimate the actual value of the inflation the next year by clicking once on the punctuated line." Instructions for those in the interval forecasting group were: "Below is a series of inflation rates for one country. WHAT WILL HAPPEN NEXT? Please make your 90% prediction interval. (90% prediction intervals correspond to the interval in which future observations will fall, with a 90% probability.) Click twice on the punctuated line at the end of the graph to show the upper and lower boundary of this 90% interval." Finally, those in the density forecasting group were instructed as

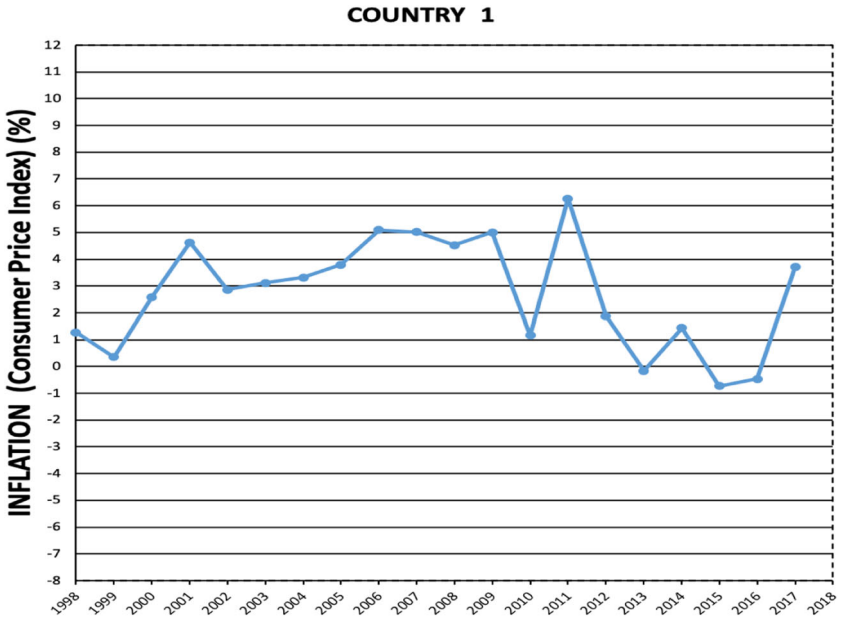


FIGURE 1 Experiment 1: Time series showing 20 years of annual inflation figures from Country 1

follows: "Below is a series of inflation rates for one country. WHAT WILL HAPPEN NEXT? Please allocate £100 to the 20 bins appearing on the screen. Money allocation should be higher in the bins where you believe there is a greater probability for actual inflation next year. To allocate all £100, please enter your bets to each of the bins at the end of the graph." These instructions appeared above each of the 10 graphs for which participants had to make forecasts. Examples of the screens in the three conditions are shown in Figure 2.

2.1.2 | Results

The central tendency of forecasts that each person made in the three conditions was first extracted. For those in the point forecasting group, this was simply the point forecasts that they made. For those in the interval forecasting group, it was the mid-point of the interval bounds that they provided. For those in the density forecasting group, we used the reported bin probabilities to fit an underlying parametric density and then extracted the underlying forecast density mean from this. This approach, developed by Engelberg et al. (2009), assumes that probabilistic beliefs are unimodal and that a participant's distribution can be specified as a member of the generalized Beta family. However, when a forecaster fills in values for only two of the 20 bins, it is only possible to specify the mean of distribution when the two bins are adjacent. In our experiment, 10 participants failed to do this. As a result, our sample for the density forecasting group was reduced.

Once we had extracted the mean value of forecasts on each trial for each participant in each condition, we carried out a two-way mixed analyses of variance (ANOVA) on these values using forecast type (point forecasting, interval forecasting, and density forecasting) as a between-participant factor and trial number (1–10) as a within-participant factor. This analysis showed no significant main effects or interactions. Thus, we obtained no evidence inconsistent with H_1: even with a between-participants design that excluded the possibility of anchoring effects, there was no suggestion that different forecasting methods produced different estimates of the central value of inflation. This replicates and reinforces Bruine de Bruin, Manski, et al.'s (2011) conclusions.

It is clear from the upper panel of Figure 3 that people overestimated inflation: a one-sample t test showed that ME was significantly positive, t (128) = 22.67, $p < .001$. This finding is consistent with H_2. A one-way ANOVA on ME using forecast type as a between-participant factor revealed no significant main or interactive effects. Thus, we failed to obtain evidence inconsistent with H_3.

A one-way ANOVA on VE revealed an effect of forecast type, F (2, 126) = 12.43, $p < .001$, generalized eta squared (ges) = 0.1648.[6] Post hoc analyses revealed significant differences between the point and density forecast ($p < .001$) and between the point forecast and the interval forecast ($p = .002$) but no significant difference between the interval forecast and the density forecast (Figure 3, middle panel).

Finally, a one-way ANOVA on RMSE revealed an effect of forecast type, F (2, 126) = 6.56, $p = .002$, ges = 0.0943. Post hoc analyses revealed significant differences between the point and density forecast ($p = .01$) and between the point forecast and the interval forecast ($p = .007$) but no significant difference between the interval forecast and the density forecast. This is shown in the lower panel of Figure 3.

Our hypotheses do not concern the relative quality of uncertainty estimation in interval and density forecasting. However, in selecting between those two types of forecasting, users may wish to take this issue into account. Hence, we present an analysis of it in Appendix A.

2.1.3 | Discussion

Despite using a well-powered between-participant design to eliminate anchoring and other context effects, we obtained no evidence

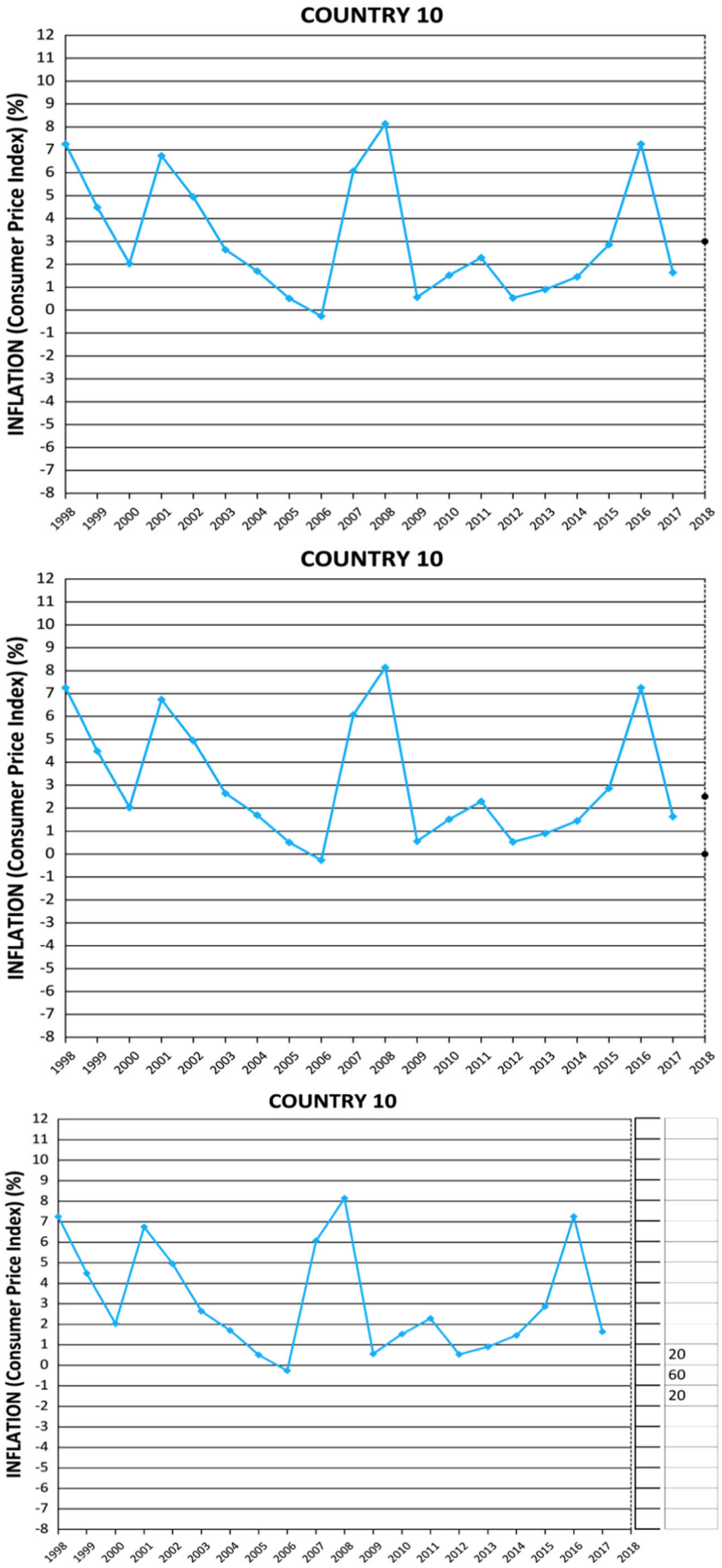


FIGURE 2 Experiment 1: Examples of point forecasting (upper panel), interval forecasting (middle panel), and density forecasting (lower panel)

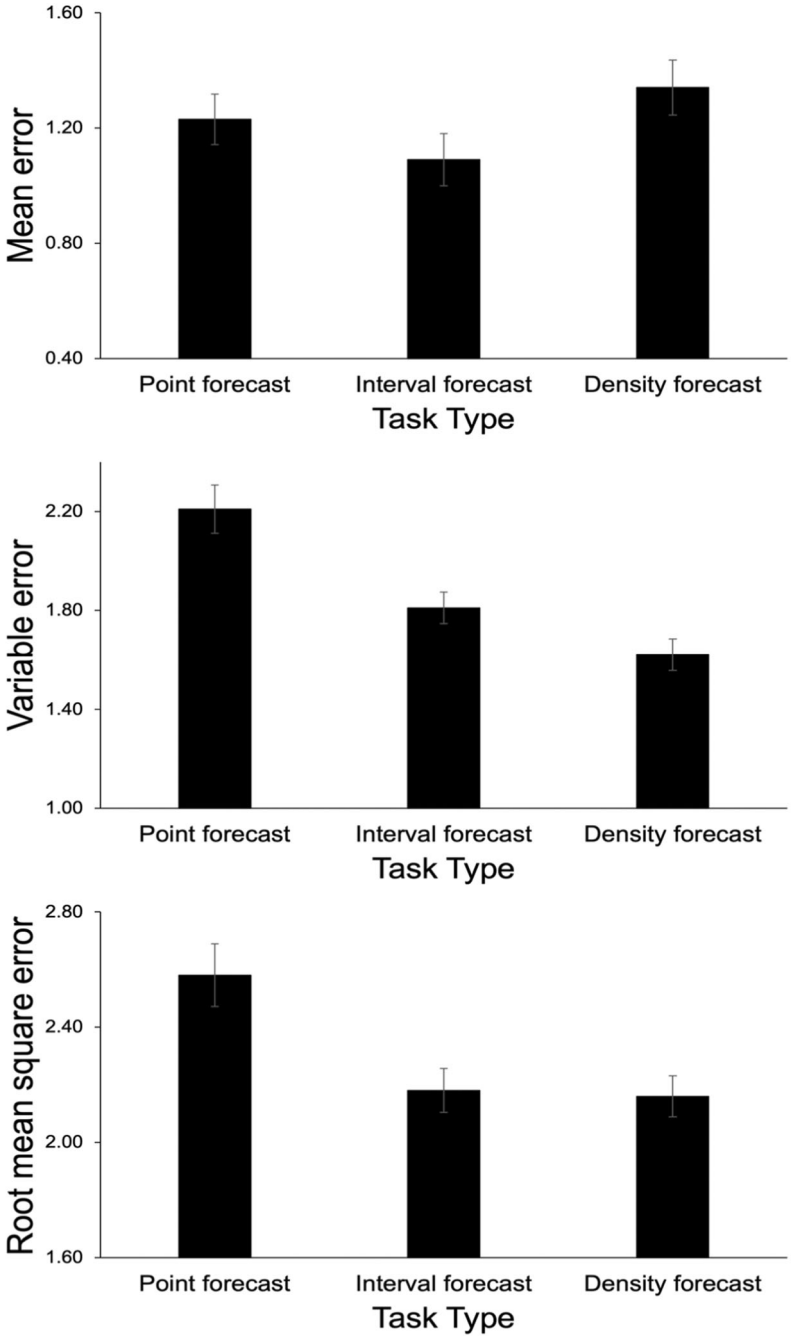


FIGURE 3 Experiment 1: Bar chart showing mean error, variable error, and root mean-squared error scores (with standard error bars) in each forecasting task

against H_1. If there is any effect of type of forecast on estimates of central tendency produced by different types of forecast, it must be small. Thus, data from our between-participant design replicates the finding that Bruine de Bruin, Manski, et al. (2011) obtained in their within-participant design. This implies that context effects did not influence their result.

Our finding that people overestimated inflation rate is consistent with H_2 and replicates findings from previous studies (e.g., Bruine de Bruin, van der Klaauw, et al., 2011; Bryan & Venkatu, 2001a, 2001b; Georganas et al., 2014). It appears that people expect inflation to be higher than it turns out to be.

Given the lack of evidence against H_1, it is not surprising that there was also no evidence suggesting that bias in inflation forecasts depends on the type of forecast made (H_3).

We expected that estimates of central tendency derived from point forecasts would be noisier and less accurate than those derived from interval and density forecasts and that those derived from interval forecasts would be noisier and less accurate than those derived from density forecasts (H_4). This received partial support: point forecasts were indeed noisier and less accurate than estimates of central tendency derived from interval and density forecasts but estimates of central tendency derived from interval forecasts were

not noisier and less accurate than those derived from density forecasts. Our hypothesis was based on the "wisdom of the inner crowd" effect (Herzog & Hertwig, 2009, 2014; Van Dolder & van den Assem, 2018; Vul & Pashler, 2008): the average of a number of judgments from a single person produces a more accurate estimate than a single judgment from that person. We argued that the mean value of the forecast distribution is estimated just once in point forecasting, twice (albeit implicitly) in interval forecasting, and three or more times (albeit implicitly and depending on the number of bins filled) in density forecasting.

Why did we fail to obtain a difference between the VE and RMSE values associated with interval and density forecasting? First, the benefit gained from averaging more judgments decreases with the number of judgments already averaged. Referring to the increases in accuracy obtained by averaging judgments from more advisors, Budescu and Yu (2007, p. 154) point out that "The accuracy of the average opinion increases monotonically as a function of the number of advisors but at a diminishing rate that depends on the inter-judge correlation." Thus, it could be that most of the gain to be obtained by aggregating judgments is associated with increasing the number of judgments from one (point forecasting) to two (interval forecasting) and that little extra benefit is obtained by increasing the number of judgments beyond two (density forecasting).

Second, the difficulties that people are reported to have in making density forecasts (O'Hagan et al., 2006) may have increased the random noise in their responses and thereby canceled out any benefit derived from repeatedly making implicit judgments of the central tendency of the distribution. In contrast, providing interval forecasts is a simple task and so the benefit derived from making an estimate of the central tendency twice would not be diluted by judgment noise associated with performing a difficult task.

2.2 | Experiment 2

Our first experiment showed that, relative to point forecasting, the gain in accuracy from using interval forecasts was as great as the gain in accuracy from using density forecasts. Furthermore, unlike point forecasts, interval forecasts provide survey users with some information about respondents' estimates of the aleatory uncertainty[7] in the system responsible for generating inflation. This should be important for predicting consumers' behavior: they are less likely to act on less certain inflation forecasts.

Thus, the mid-point of the bounds of an interval forecast provides a more accurate estimate of the expected value of inflation than a point forecast and interval forecasts also provide uncertainty information. Furthermore, they are no less accurate than density forecasts and much simpler to produce. Their only drawback is that the uncertainty information that they provide is not as detailed as and not as accurate as that produced by density forecasts. However, if survey users do not require uncertainty information that is as detailed and accurate as that obtained from density forecasts, interval forecasts would provide advantages over point forecasts without incurring the disadvantages of density forecasts.

Surveys ask participants many questions: responses to the earlier questions may influence how later ones are answered (Niu & Harvey, 2021). Although the last experiment indicated that such context effects did not influence Bruine de Bruin, Manski, et al.'s (2011) finding that the mean value of central forecast is unaffected by the type of forecast made, it is possible that such effects may differentially influence the noisiness (VE) and accuracy (RMSE) of different types of forecast. For example, the accuracy advantage of interval forecasts may vanish when they are made after point forecasts. Thus, in this experiment we ask whether the accuracy advantage of interval forecasts is preserved in a within-participant design. Using this design, we address the same hypotheses as before. We also examine two other issues.

If context effects do influence the accuracy of different types of forecast (without influencing their mean value), the strength of that influence may be affected by the order in which the different types of forecast are made. Explicit point forecasts may provide stronger anchors for interval forecasts than the (implicit) central values of interval forecasts provide for point forecasts. If they do, point forecasts would reduce judgment noise (VE) in interval forecasts that follow them more than interval forecasts would reduce judgment noise in point forecasts that follow them. Thus, the order in which the two types of forecasts are made may affect the noisiness of interval forecasts more than that of point forecasts. With this in mind, we test the following hypothesis.

$\mathbf{H_6}$: $|VE_{\text{Interval second}} - VE_{\text{Interval first}}| > |VE_{\text{Point second}} - VE_{\text{Point first}}|$.

In this experiment, we also measure people's confidence in the judgments. We asked them to assess the likelihood (0%–100%) that their point forecast or their interval bounds were within 10% either way of their true values. In other words, we asked them to assess their epistemic uncertainty in their own judgments. As their interval forecast (but not their point forecast) provided their estimate of the aleatory uncertainty in the inflation figures, this allowed us to examine how aleatory and epistemic uncertainty are related (Tannenbaum et al., 2017). People may be more confident in judgments in which they have been allowed to express their uncertainty (interval forecasts) than in those in which they have not (point forecasts).

$\mathbf{H_7}$: $\text{Confidence}_{\text{Point Forecasts}} < \text{Confidence}_{\text{Interval Forecasts}}$.

2.2.1 | Method

Two groups of participants made both point forecasts and interval forecasts: one of those groups made point forecasts followed by interval forecasts and the other group made them in the reverse order.

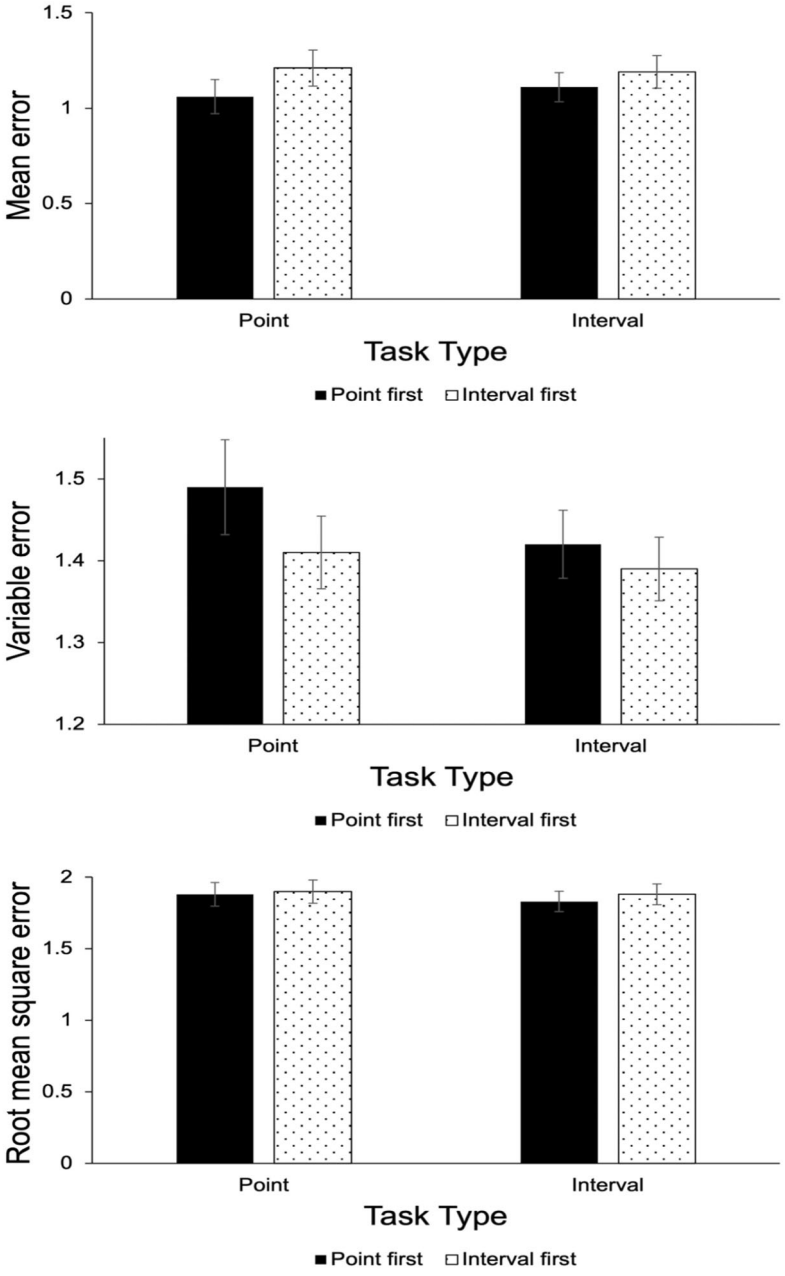


FIGURE 4 Experiment 2: Bar charts showing mean error, variable error, and root mean-squared error scores (with standard error bars) in each condition

Participants

One hundred and one participants (65 males, 39 females) with a mean age of 29 years (SD = 11 years) took part in the web-based study. They were recruited from the online participant recruitment platform, www.Prolific.com, between July 16, 2020 and August 13, 2020 and paid £1.10 for their participation.

Stimulus materials

Participants made forecasts from 10 graphs, each showing 20 years (2000 to 2019) of real inflation data from 10 countries that were extracted from the World Bank website. For each one, they made both a point forecast and an interval forecast for the 2020 inflation rate. The order of these judgments varied between participants. Both types of forecast were made in the same way as in Experiment 1.

After each forecast, participants expressed their confidence in it by moving a slider that ranged between 0% and 100%. For the point forecast, participants were first asked "How confident are you about the point forecast you just made?" and then moved their slider along a scale that was labeled at the left end "0%—My estimate is **definitely not accurate** to within 10% either way of the correct value," in the middle "50%—My estimate is **as likely to be** within 10% of the true value as it is not to be within that range," and at the right end "100%—My estimate is **definitely accurate** to within 10% either way of the true value." The chosen position on the slider was made numerically explicit with a message posted below it: for example, "You are 60% confident about your forecast." To reduce anchoring effects, the starting position of the slider was randomized from the 21 possibilities that were 5% apart (0%, 5%, 10%, … 100%).

For the interval forecast, participants were first asked "How confident are you about the interval boundaries that you set?" and then moved their slider along a scale that was labeled at the left end "0%—My interval boundaries are **definitely not accurate** to within 10% either way of the correct value," in the middle "50%—My interval boundaries are **as likely to be** within 10% of the true value as it is not to be within that range," and at the right end "100%—My interval boundaries are **definitely accurate** to within 10% either way of the true value." The chosen position on the slider was made numerically explicit with a message posted below it: for example, "You are 60% confident about your interval forecast." To reduce anchoring effects, the starting position of the slider was randomized from the 21 possibilities that were 5% apart (0%, 5%, 10%, … 100%).

For the same reasons as before, inflation data series were displayed in the central part of the *y*-axis scale, which ranged from 10% to −6%. The final points for all 10 countries were in the middle of that scale, ranging between 0.08% and 2.90%. The 10 series were broadly comparable, with a mean inflation of 2.05% (SD = 1.75%).

Design

Forecast type (point forecast and interval forecast) was varied within participants. The order of these tasks was varied between participants, who were randomly allocated to a point-then-interval group ($N = 43$) or to an interval-then-point group ($N = 44$). The presentation order of the 10 graphs in each task was individually randomized for each participant. The identities of the 10 real countries were anonymized by labeling them with numbers: for example, "Country 3 of 10."

Procedure

The procedure and task instructions for point and interval forecasts were the same as those described for Experiment 1.

2.2.2 | Results

Forecasts for each country were compared with the actual 2020 inflation rates that were extracted from the World Bank website.

Respondents were excluded from the analysis if any of their forecasts (point forecasts or mid-point of the interval forecasts) were beyond three standard deviations of the mean forecast for a particular country and forecast type. This led to a sample for analysis of 87 people (56 males, 31 females) with a mean age of 28 years (SD = 9 years).

Forecasts

After extracting the mean value of forecasts on each trial for each participant in each condition, we carried out a three-way ANOVA on these values using task order (point forecasting first, interval forecasting first) as a between-participant factor and forecast type (point forecasting, interval forecasting) and trial number (1–10) as within-participant factors. This analysis showed no significant main effects or interactions. Thus, we again failed to obtain evidence inconsistent with H_1 but, this time, in a within-participants design of the sort used by Bruine de Bruin, Manski, et al. (2011).

Again, people systematically overestimated inflation (Figure 4, upper panel): a one-sample *t* test showed that ME was significantly positive, $t(86) = 19.14$, $p < .001$, consistent with H_2. However, a two-way mixed ANOVA on ME with task order as a between-participant variable and forecast type as a within-participant one revealed no significant effects. As in Experiment 1, we failed to obtain evidence inconsistent with H_3.

A two-way mixed ANOVA on VE (Figure 4, middle panel) using the same factors as before revealed only a main effect of forecast type, $F(1, 85) = 4.09$, $p = .046$, $ges = 0.0058$. This provides further evidence consistent with H_4. However, there was no evidence for the interaction predicted by H_6: the relative noisiness of point and interval forecasts was not affected by the order in which they were made.

Finally, a two-way mixed ANOVA on RMSE using the same factors as before yielded no significant effects. Despite forecast type significantly affecting VE, this did not feed through to producing a correspondingly significant effect on RMSE (Figure 4, lower panel). Presumably, this was because any such effect was overwhelmed by the influence of ME on RMSE.[8]

Confidence in forecasts

For each confidence judgment (e.g., there is a 60% chance of my point forecast is within 10% of the true value), we set an outcome index, *d*, at 1.00 when the event occurred (the forecast was within 10% of the true value) and at 0.00 when the event did not occur (the forecast was not within 10% of the true value). On each trial, the difference $(j - d)$ between the judgment, *j*, expressed as a probability rather than as a percentage, and the outcome index, *d*, then provides a measure of the *bias* in the judgment: higher mean values of this difference indicate greater overconfidence. The square of the bias, $(j - d)^2$, is known as the probability score: lower mean values of the probability score indicate a greater ability to assign appropriate probabilities (Yates, 1990, 1994). Here we subtract the probability score from 1.00 so that higher values indicate better calibration: we term this the calibration score.

ANOVAs using the same three factors used for the analysis of forecasts indicated that forecast type had an effect on level of confidence, $F(1, 85) = 27.83$, $p < .0001$, $ges = 0.0299$, on bias, $F(1, 85) = 44.78$, $p < .001$, $ges = 0.0433$, and on the calibration score, $F(1, 85) = 36.10$, $p < .001$, $ges = 0.0354$: people were more confident in their interval forecasts than in their point forecasts, more overconfident in them, and less able to judge how likely they were to be accurate (Figure 5).

2.2.3 | Discussion

The switch from a between-participants to a within-participants design had little effect on the nature of the findings that we obtained. Again, there was no significant difference between mean values of point forecasts and the mean values of the mid-point of interval forecasts (H_1). Furthermore, though there was significant

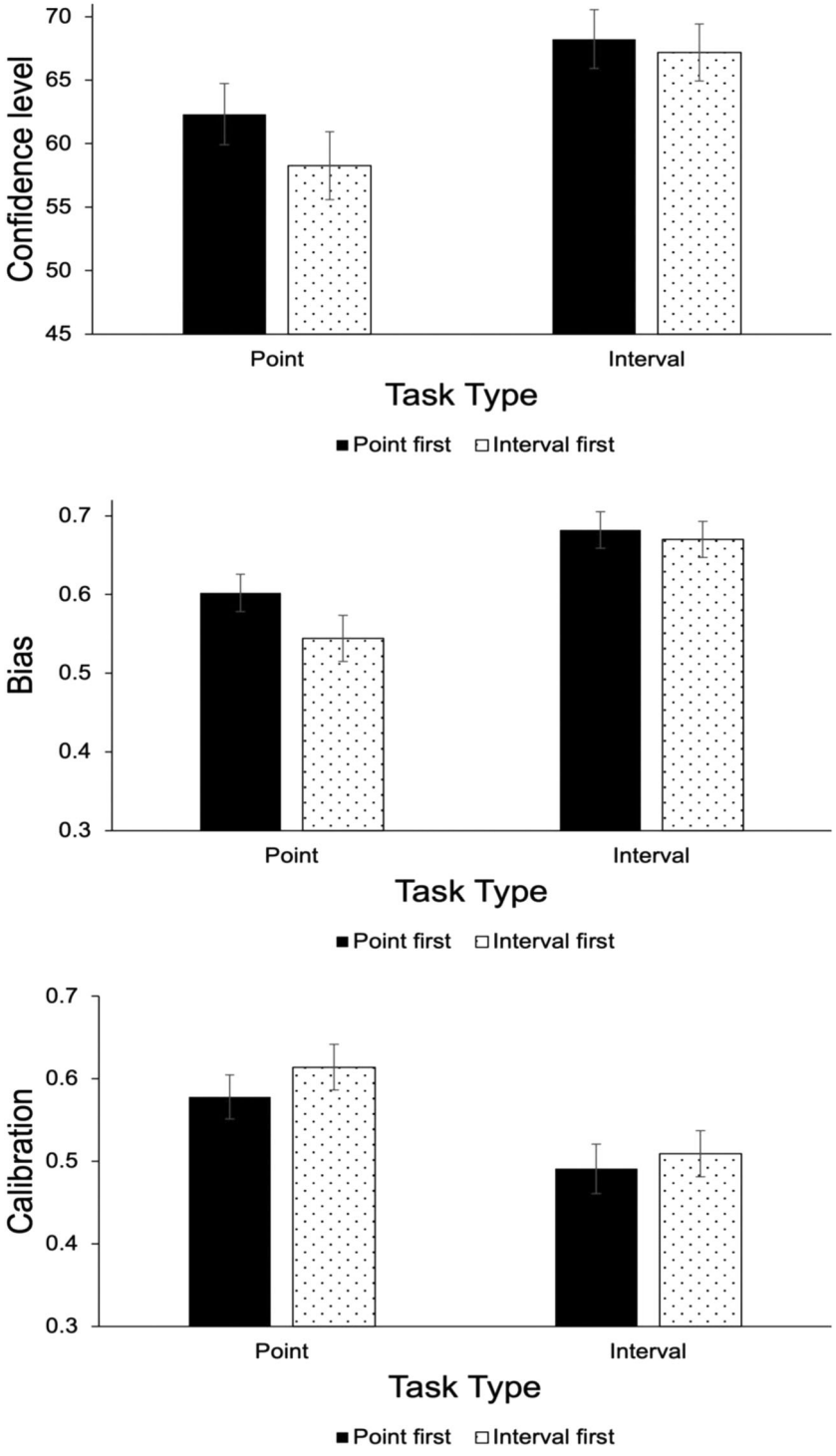


FIGURE 5 Experiment 2: Bar charts showing mean confidence levels, bias levels, and calibration scores (with standard error bars) in each condition

over-forecasting of inflation (H_2), the size of this bias did not depend on the type of forecast made.

It would, however, be wrong to assume that the change in experimental design had no effect. In our previous between-participant experiment, the mean value of RMSE was 2.58 in the point forecasting condition and 2.18 in the interval forecasting condition: the former was 18% higher than the latter. In contrast, in this experiment, the corresponding RMSE values for point and interval forecasting were 1.89 and 1.85, respectively: the former was only 2% higher than the latter. In other words, the use of a within-participant rather than a between-participant design reduced the percentage difference in RMSE between point and interval forecasting by nine-tenths. This is consistent with anchoring having an effect in reducing the difference in the accuracy of central forecasts derived from different methods of elicitation. However, in contrast, to claim of H_6, this anchoring effect was not asymmetrical: initial point forecasts acted as mental anchors for the central value of later interval forecasts to the same extent that the central value of initial interval forecasts acted as

mental anchors for later point forecasts. As a result, there was no effect on the order in which the two types of forecasts were made.

Although the mean of the bounds of interval forecasts provided less variable estimates of future inflation than point forecasts, people were more overconfident in their ability to estimate those bounds than in their ability to make point forecasts. It is possible that allowing people to express uncertainty in the forecasts (i.e., making interval forecasts) reduces their concern about being wrong and that this, in turn, raises their confidence in their accuracy. Whatever the mechanism, higher confidence in interval judgments could have implications for survey design and the use of survey data.

We suggested that the mean of the bounds of an interval forecast provides a more accurate central forecast than a point forecast because people make two separate (though implicit) point forecasts when estimating intervals. Because of the "wisdom of the inner crowd" effect (Herzog & Hertwig, 2009, 2014; Vul & Pashler, 2008), this acts to cancel out judgment noise (Kahneman et al., 2021) and so increases accuracy. This can explain the difference between these two types of forecasting in both a between-participants design (Figure 3) and a within-participants design (Figure 4). In the next experiment, we test predictions arising from this account.

2.3 | Experiment 3

If the "wisdom of the inner crowd effect" is responsible for the mid-point of the bounds of an interval forecast providing a more accurate estimate of the expected value of inflation than a point forecast, then asking people to make a point forecast from the same data on two separate occasions and taking the average of those judgments should result in less noisy and, hence, more accurate estimates of inflation than either one of those judgments separately. In other words, the VE and RMSE of the average of the two point forecasts should be less than the average VE and average RMSE of the two separate forecasts.

H_8: VE of average of two point forecasts < Average VE of two point forecasts.

H_9: RMSE of average of two point forecasts < Average RMSE of two point forecasts.

The "wisdom of the inner crowd" effect may or may not be sufficient to explain the difference in accuracy of point and interval forecasts. If it is sufficient,

H_{10}: VE of average of two point forecasts = VE of the mid-point of interval forecast bounds.

H_{11}: RMSE of average of two point forecasts = RMSE of the mid-point of interval forecast bounds.

H_{12}: Average VE of two point forecasts > VE of the mid-point of interval forecast bounds.

H_{13}: Average RMSE of two point forecasts > RMSE of the mid-point of interval forecast bounds.

The "wisdom of the inner crowd" effect is assumed to arise because averaging two or more judgments cancels out some of the noise in the judgments (Kahneman et al., 2021). Thus, the effect should be greater when single judgments contain more noise. If separate judgments were noise-free, we would expect no reduction in RMSE after averaging them; if separate judgments were very noisy, we would expect averaging them to produce a large reduction in both VE and RMSE. People's time-series forecasts contain more noise when the data series on which they are based contain more noise (Harvey, 1995; Harvey et al., 1997). Thus, we expect the "wisdom of the inner crowd" effect to be greater when people make forecasts from more volatile series.

H_{14}: The effect identified in H_8 and H_9 will be greater with noisier inflation series.

H_{15}: The effect identified in H_{12} and H_{13} will be greater with noisier inflation series.

2.3.1 | Method

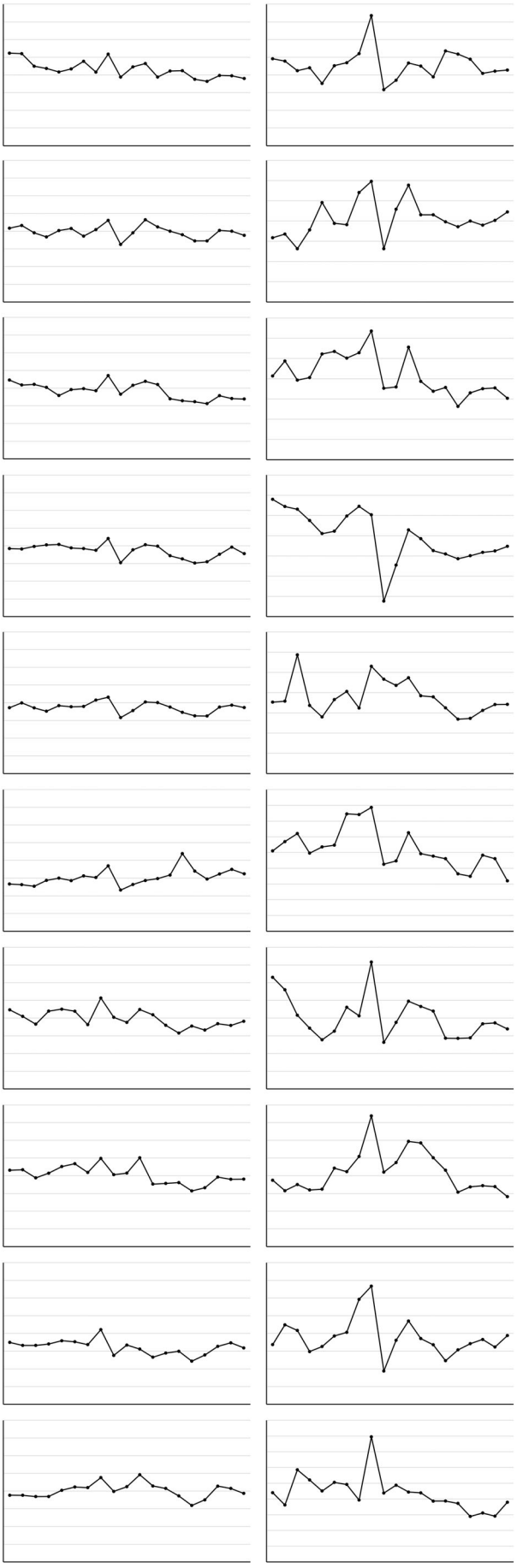
To test these hypotheses, we ran two groups of participants. The first group made interval forecasts of inflation for 20 countries. The second group made point forecasts of inflation for those 20 countries and then made point forecasts for those same countries again (in the same order).

Participants

One hundred and one participants (66 males, 35 females) with a mean age of 26 years (SD = 10 years) took part in the web-based study. They were recruited from the online participant recruitment platform, www.Prolific.com, between November 28, 2020 and January 19, 2021 and paid £1.00 for their participation.

Stimulus materials

Stimulus graphs depicting 20 years (2000–2019) of real inflation data from 20 countries extracted from the annual CPI data set provided by the World Bank. The series were displayed in the central part of the *y*-axis scale and ranged from 10% to −6%. The final points for all 20 countries were in the middle of that scale, ranging between −0.36% and 2.90%. For the 10 low volatility series, the mean inflation was 1.58% (SD = 1.11%); for the 10 high volatility series, the mean inflation was 2.45% (SD = 2.28%). Mean levels of variance were 0.70 (SD = 0.18) for the 10 low volatility series and 4.50 (SD = 1.20) for the 10 high volatility series. These were significantly different, t (18) = 8.94, p < .001. The two sets of series are shown in Figure 6. It is clear that the greater volatility in the second set of countries occurred mainly around the time of the worldwide financial crisis (2008–2011). As before, countries were



not explicitly named in the experiment: they were referred to by number.

Design

Forecast type was a between-participant variable: one group made interval forecasts from the 20 series; the other group made point forecasts from those series and then made another set of point forecasts from those same 20 series in the same order as before. Series volatility was a within-participant variable: forecasts were made for 10 low volatility and 10 high volatility series. The order of the 20 series was randomized separately for each participant. In the point forecast condition, each participant received the series in the same order in the first and second blocks of forecasts. There was no explicit separation of these blocks: as far as participants were concerned, they made 40 forecasts in a single block of 40 series.

Procedure

As before, participants saw an information screen and responded to a consent screen before receiving a simple definition of inflation (Consumer Price Index) and being given their instructions. Instructions for the point forecasting and interval forecasting groups were as described for the previous experiments. An example picture of an interval forecast or a point forecast (upper two panels of Figure 2) was provided before the start of the formal task. At the end of the experiment, participants answered the same demographical questions as before.

2.3.2 | Results

Two participants were excluded because of missing values in their data. Others were excluded from the analyses using the same criteria as specified for Experiment 2. As a result, data were analyzed from 85 participants (54 males, 31 females) with a mean age of 27 years (SD = 10 years). There were 45 participants in point forecast condition and 40 participants in interval forecast condition.

We calculated ME, VE, and RMSE scores derived (a) from the average value of the two point forecasts, (b) separately for each point forecast and then averaged, and (c) from the mid-point (average) of the bounds of the interval forecast. These different types of error scores are shown in Figure 7 for low volatility series (left bar of each pair) and high volatility series (right bar of each pair). Inspection suggests that volatility affects the ME for point, but not interval forecasts, VE is comparable for the average value of the two point forecasts and the mid-point of the interval forecasts, but that both of

FIGURE 6 Experiment 3: The 20 inflation series each showing 20 years of historical inflation data between 1998 and 2017 and ranging between −6% and 10%. Low volatility series are shown on the left and high volatility series on the right. In this figure, axis labels and numbering have been excluded for clarity but they were included in the experimental materials

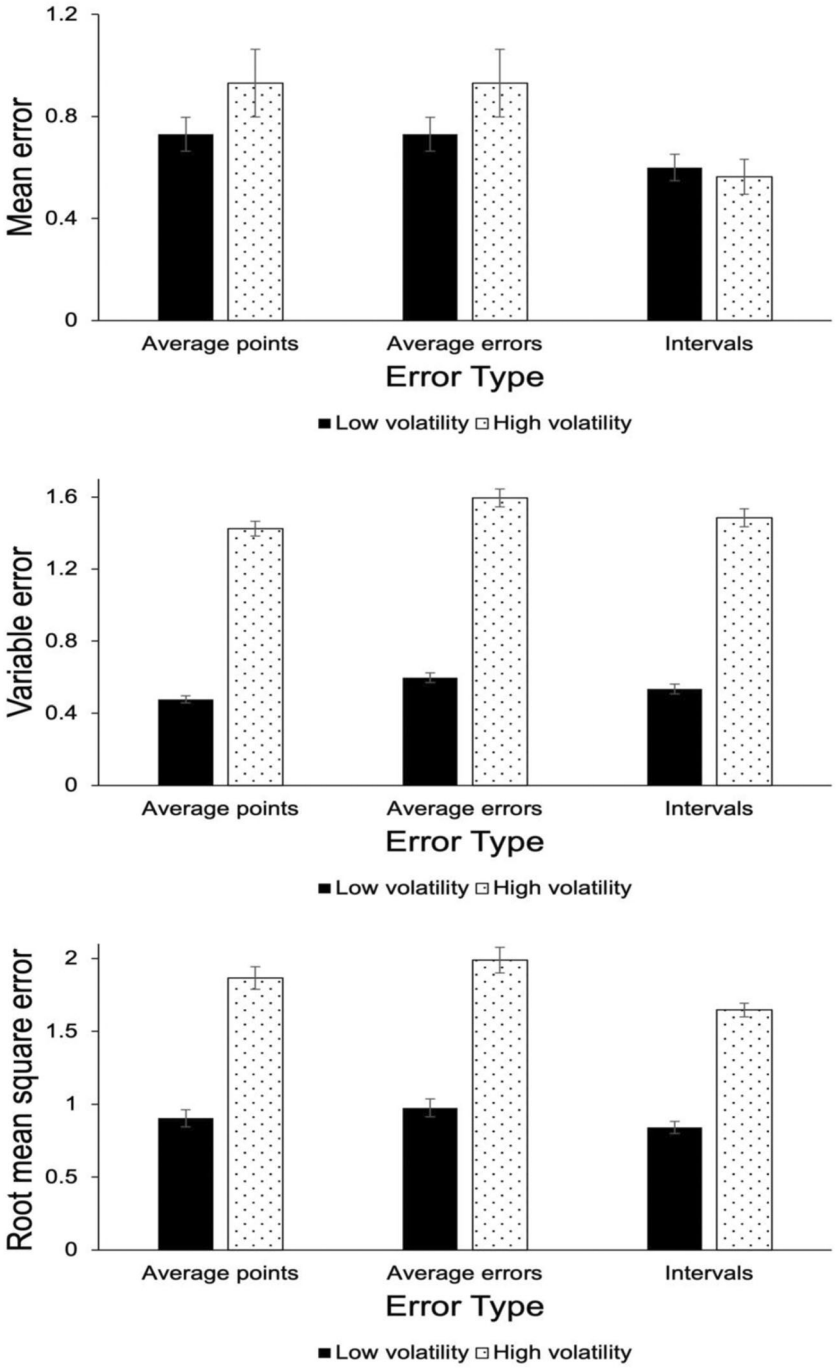


FIGURE 7 Experiment 3: ME (Upper panel), VE (Middle panel), and RMSE (Lower panel) scores for inflation forecasts made from low and high volatility series. Within each panel, the scores are based on the average value of the two point forecasts (Left), the average error of the two point forecasts (Centre), and the mid-point of the interval forecasts (Right)

these are lower than the average VE of the two point forecasts, and the pattern of RMSE scores closely reflects the combined values of the ME and VE scores.

To test H_8, H_9, and H_{14}, we carried out ANOVAs on VE and RMSE values derived from the point forecasting group data with type of point forecast error (average error of the two separate forecasts vs. error of the average of the two forecasts) and volatility (high/low) as within-participant variables. The ANOVA on VE scores revealed main effects of type of forecast error, F (1, 44) = 52.06, p < .001, *ges* = 0.0831, and volatility, F (1, 44) = 447.95, p < .001, *ges* = 0.8038, and a significant interaction between these factors, F (1, 44) = 4.12, p < .05, *ges* = 0.0024. Follow-up analyses revealed that the simple effect of forecast type error was significant for both high volatility series, F (1, 44) = 35.39, p < .001, *ges* = 0.0719, and low volatility series, F (1, 44) = 52.82, p < .001, *ges* = 0.1332, and that the simple effect of volatility was significant for both the average VE of the two point forecasts, F (1, 44) = 401.36, p < .001, *ges* = 0.7786, and the VE of the average of the two point forecasts, F (1, 44) = 444.70, p < .001, *ges* = 0.8338.

The ANOVA on RMSE scores revealed main effects of type of forecast error, F (1, 44) = 63.48, p < .001, *ges* = 0.0101, and volatility, F (1, 44) = 294.62, p < .001, *ges* = 0.5163, and a significant interaction

between these factors, F (1, 44) = 11.54, p = .001, ges = 0.0007. Follow-up analyses revealed that the simple effect of forecast type error was significant for both high volatility series, F (1, 44) = 48.76, p < .001, ges = 0.0122, and low volatility series, F (1, 44) = 51.07, p < .001, ges = 0.0081, and that the simple effect of volatility was significant for both the average VE of the two point forecasts, F (1, 44) = 268.88, p < .001, ges = 0.5064, and the VE of the average of the two point forecasts, F (1, 44) = 317.58, p < .001, ges = 0.5280. In summary, these results are consistent with H_8, H_9, and H_{14}.[9]

To test H_{10} and H_{11}, we carried out mixed ANOVAs on the VE and RMSE values with type of forecast error (derived from the average of two point forecasts vs. derived from the mid-point of the bounds of interval forecasts) as a between-participants factor and with volatility (high/low) as a within-participants factor. The ANOVA on VE revealed only a main effect of volatility, F (1, 83) = 786.17, p < .001, ges = 0.8097. Thus, we obtained no evidence inconsistent with H_9.

However, the ANOVA on RMSE revealed a main effect of volatility, F (1, 83) = 545.50, p < .001, ges = 0.5708, and an interaction between volatility and type of forecast error, F (1, 83) = 4.30, p < .04, ges = 0.0104. Further analysis showed that the simple effect of volatility was significant both for the RMSE derived from the point forecasts, F (1, 44) = 317.58, p < .001, ges = 0.5280, and for the RMSE derived from the interval forecasts, F (1, 39) = 235.62, p < .001, ges = 0.6749, but that the simple effect of type of forecast error was significant only for the high volatility series, F (1, 83) = 5.54, p = .02, ges = 0.0625.

The significance of this interaction for RMSE, but not for VE, strongly suggests that it arose because the ME for the average of the two point forecasts was affected much more by higher volatility than the ME associated with the mid-point of the bounds of the interval forecast. An ANOVA on the ME scores using the same factors as before confirmed this. It revealed a main effect of type of forecast error, F (1, 83) = 4.74, p = .03, ges = 0.0459, and an interaction between that variable and volatility, F (1, 83) = 5.70, p = .02, ges = 0.0107. ME values for point forecasts, but not for interval forecasts, were higher when series were more volatile. Thus, the simple effect of volatility was significant only for the point forecasts, F (1, 44) = 5.82, p = .02, ges = 0.0204, and the simple effect of type of forecast error was significant only for high volatile series, F (1, 83) = 5.66, p = .02. ges = 0.0638. In summary, we found no evidence inconsistent with H_{10} but we did obtain evidence inconsistent with H_{11} because of this unexpected effect of forecast type on ME, one of the contributors to RMSE.

Finally, to test H_{12}, H_{13}, and H_{15}, we carried out mixed ANOVAs on the VE and RMSE values with the type of forecast error (the average of the errors calculated separately for each point forecasts vs. the error derived from the mid-point of the bounds of interval forecasts) as a between-participants factor and with volatility (high/low) as a within-participants factor. The analysis of VE revealed main effects of type of forecast error, F (1, 83) = 3.87, p = .05, ges = 0.0274, and volatility, F (1, 83) = 744.03, p < .001, ges = 0.7805, but no interaction between these factors.

The analysis of RMSE revealed main effects of type of forecast error, F (1, 83) = 8.75, p = .004, ges = 0.0772, and volatility, F (1, 83) = 490.41, p < .001, ges = 0.5491, together with an interaction between these variables, F (1, 83) = 6.38, p = .01, ges = 0.0156. Further analysis revealed a simple effect of type of forecast error only for high volatility series, F (1, 83) = 10.97, p = .001, ges = 0.1168, and a simple effect of volatility for both point forecast error, F (1, 44) = 268.88, p < .001, ges = 0.5064, and interval forecast error, F (1, 39) = 235.62, p < .001, ges = 0.6749. Again, this interaction was observed for RMSE but not for VE because of the unexpected effect of volatility on the ME of point forecasts but not on the ME of interval forecasts. (Because ME is the same whether it is calculated as the average ME of the two separate point forecasts or as the ME of the average of the two point forecasts, the analysis of ME reported in the last paragraph applies here too.) In summary, these results are consistent with H_{12} and H_{13}, but only partially consistent with H_{15}.

2.3.3 | Discussion

Comparison of VE and RMSE scores for the two ways of averaging point forecasts revealed results that we expected. Error scores based on the average of the two forecasts were significantly lower than averages of the error scores calculated for each forecast separately. This was expected on the basis of a "wisdom of the inner crowd" effect. Furthermore, if this effect reduces VE by some proportion (e.g., 20%), then the absolute size of the reduction should be greater when VE is higher. VE is higher when forecasts are made from noisier series (Harvey, 1995). Hence, we expected the "wisdom of the inner crowd" effect to be greater with the noisier series; the interactions between series volatility and forecast type for VE and RMSE provide evidence that it was indeed greater for noisier series.

Turning now to the comparisons between VE based on the mid-point of the bounds of the interval forecasts and the two values of VE based on the different methods of obtaining VE from point forecasts, we found that there was no evidence of a difference between the VE values derived from the interval forecasts and the VE values derived from the average of the two point forecasts. There was, however, a significant difference between the VE values derived from the interval forecasts and the average of the VE values calculated separately for each of the two point forecasts. Taken together, these results are consistent with the "wisdom of the inner crowd" effect being sufficient to explain why VE values derived from the mid-point of the bounds of interval forecasts are lower than those derived from single point forecasts (Experiments 1 and 2).

Consistent with this, there was a main effect of volatility but no interaction between volatility and type of forecast when the VE values derived from the mid-points of the bounds of interval forecasts were compared with the VE values derived from the average of the two point forecasts: no interaction was expected because there was no difference in the size of those VE values. However, we had expected the corresponding interaction to be significant when VE values derived from the mid-points of the bounds of interval forecasts

were compared with the average of the VE scores calculated separately for each of the two point forecasts. In fact, this interaction did not attain significance.

There was one other way in which the results did not turn out in the manner we had predicted. In developing our hypotheses, we had expected that the results from the analyses of RMSE values would broadly reflect those obtained from the analyses of VE scores. Our expectations were based on the assumption that ME scores would not be influenced by volatility or by whether they were derived from point forecasts or from interval forecasts. In fact, volatility increased not only VE (which we had expected) but also ME (which we had not expected). Furthermore, this effect of volatility on ME was restricted to ME values derived from point forecasts; it was not present for ME values derived from interval forecasts. Because of these differential effects of volatility on different error sources, results from analyses of RMSE did not reflect those from our analyses of VE in the way we had expected. In particular, analysis of RMSE scores derived from the mid-points of the bounds of interval forecasts and those derived from the averages of the two point forecasts revealed an interaction between volatility and error source (whereas analysis of the corresponding VE scores did not). Similarly, analysis of RMSE scores derived from the mid-points of the bounds of interval forecasts and the average RMSE scores calculated for each point forecast separately also revealed an interaction between volatility and error source (whereas the corresponding VE scores did not show that interaction between even though it was expected).

Why did higher volatility increase ME scores that are derived from point forecasts? There have been many demonstrations that lay people tend to expect inflation to be higher than it turns out to be (e.g., Bruine de Bruin, van der Klaauw, et al., 2011; Bryan & Venkatu, 2001a, 2001b; Georganas et al., 2014). When series contain little noise, they are constrained in how much they can over-forecast inflation while still producing a plausible prediction. However, when series are noisy, greater overforecasting is possible because forecasts well above the statistical expectation may still be plausible if they are within the envelope provided by previous outliers. Why did higher volatility not increase ME scores that are derived from the mid-points of interval forecasts? People producing interval forecasts are likely to respond to higher volatility not by taking the opportunity of raising the mid-points of their interval forecasts but by simply widening the interval that they provide (without changing its mid-point).

3 | GENERAL DISCUSSION[10]

Domains in which judgmental forecasts are required differ in terms of the type of forecast that is seen as most appropriate. When many forecasts are needed within a limited period of time, point forecasts are likely to be preferred even though they provide users with no information about the uncertainty associated with the forecasts. Demand forecasting frequently provides an example of such a domain. However, when users need information from forecasters about the uncertainty associated with their forecasts and are willing to sacrifice the speed of forecasting to obtain this information, interval and density forecasting are more appropriate. Which of these is selected depends on how detailed the information about the uncertainty associated with the forecasts needs to be: interval forecasts provide only basic uncertainty information, whereas density forecasting provides the distributional information needed in certain domains, such as finance or meteorology.

Within this overall framework, users make decisions about the type of forecast they require by identifying where their needs lie on a trade-off between speed (or convenience) of providing forecasts and the detail those forecasts contain about the uncertainty inherent in them. The work that we have reported here demonstrates that there are additional factors that should enter into their decisions. First, estimates of the expected value for the period being forecast are subject to more judgment noise in point forecasts than in interval or density forecasts: as a result, point forecasts are more inaccurate. Second, people are more overconfident in interval forecasts than in point forecasts: they are less able to assess the likelihood that their judgments are correct. Third, in domains in which forecasts are typically biased (inflation forecasting, sales forecasting), the level of bias increases with the volatility of the data on which forecasts are based when point forecasts are made but not when interval forecasts are made. Arguably, all these factors should be taken into account when users specify the type of forecast they require.

Our results are consistent with the lower variable and overall error in interval and density forecasts than in point forecasts arising from a "wisdom of the inner crowd" effect (Herzog & Hertwig, 2009, 2014; Van Dolder & van den Assem, 2018; Vul & Pashler, 2008). We suggest that the expected value or central tendency of the variable for the period being forecast is estimated just once for point forecasts but more than once for interval and density forecasts. Consistent with this, our third experiment demonstrated that the relative disadvantage of point forecasting can be eliminated by eliciting point forecasts twice and using the average of the resulting estimates. Of course, for this to be effective, an effort must be made to ensure that forecasters do not remember the exact value of their first forecast when making their second one. We accomplished this by eliciting a set of 20 forecasts and then eliciting that set again; this procedure was sufficient to produce a significant reduction in VE, an indication that people were unable to fully remember their first forecasts when making their second ones.

3.1 | Limitations

We examined point, interval, and density forecasting. Other types of forecasting are sometimes used. For example, rather than asking forecasters to set an interval such that there is a 90% likelihood that the outcome will lie within it, it is possible to provide forecasters with a fixed interval and ask them to estimate the likelihood that the outcome will fall within it. Though this latter approach is rarely used in practice, it reduces biases in forecasts (Hansson et al., 2008). In addition, the above types of forecast are sometimes used in combination: for example, forecasters may specify a point forecast and then place an interval around it or, alternatively, they may set an

interval and then place a point within it to signify the most likely outcome within the interval. These procedural variations could influence ME or VE and might lead to modification of our conclusions. However, a "wisdom of the inner crowd" approach could be used to predict how they would do so. For example, we would expect that a point forecast made before an interval forecast would contain a greater VE than one made after it.

We discussed the trade-off between the simplicity of producing a particular type of forecast and the usefulness of the forecast. For example, density forecasts are relatively difficult to produce but they provide more detailed and more accurate information to users about the uncertainty associated with the forecast. However, while sophisticated users will easily absorb this additional detail (Armstrong, 2001; Ramos et al., 2013; Roulston & Kaplan, 2009), others may find it difficult to interpret these more complex forecasts (Fischhoff, 1994; Ramos et al., 2013; Yaniv & Foster, 1995; Yates et al., 1996). For example, Bruine de Bruin, Manski, et al. (2011) reported that respondents found density forecasts for price inflation were more difficult to appreciate and less clear than point forecasts. Here we have drawn attention to one previously ignored factor (accuracy of estimation of expected values) that should be taken into account when selecting the type of forecast that will be made. However, we have not studied how forecasters and users weigh the importance of different factors when making their choice.

3.2 | Implications

Lay surveys have consistently shown that people overestimate inflation. One explanation of this is that people are more likely to recall large price increases for specific goods (e.g., rice) because those are more salient and memorable than many smaller price increases (Bruine de Bruin, van der Klaauwm, et al., 2011). For this to be a factor, the respondents must have experienced those increases in their own country. However, we obtained consistent overestimation of inflation by people who did not even know the countries for which they were making inflation forecasts. It is not easy to reconcile this finding with a model based on personal and selective recall of large price rises. It is more consistent with people having a generally biased view of inflation, perhaps because the media devotes more coverage to possible inflation increases than inflation decreases (even when those are equally likely). This "risk amplification" via the media could cause a difference in the availability of different possible levels of future inflation that influences expectations (Tversky & Kahneman, 1974).

3.3 | Conclusions

Selecting between point, interval, and density forecasting should not be just a matter of trading-off simplicity for potential usefulness. There are other relevant factors that need to be taken into account. In particular, different types of forecasting vary in how well they are able to produce estimates of the expected values of future periods of variables of interest: these estimates differ in bias, judgment noise, overall accuracy, and the degree of overconfidence associated with them.

ACKNOWLEDGMENT

None declared.

DATA AVAILABILITY STATEMENT

Data from the experiments are available at https://osf.io/m5dux/.

ORCID

Xiaoxiao Niu http://orcid.org/0000-0002-8751-4893
Nigel Harvey http://orcid.org/0000-0002-9246-1992

ENDNOTES

[1] This assumes that interval bounds are placed symmetrically around where the point forecast would be placed. Any systematic difference between the mid-point of the interval and the position of the point forecast would produce a difference in the mean error (bias) for the two types of forecast. No such difference in bias was obtained in our experiments.

[2] Our studies do not aim to simulate inflation surveys. What we ask our participants to do is not the same as what respondents are asked to do in those surveys.

[3] Mean absolute error (MAE) provides an alternative measure of overall error. We did not use it here because, unlike RMSE, it does not decompose neatly into ME and VE. However, in almost all cases reported here, analysis of it leads to the same conclusions as those arising from the analysis of RMSE.

[4] Some but not all surveys of inflation expectations provide respondents with information about inflation in preceding periods.

[5] At the end of the experiment, participants also answered three open-ended questions about how they made their forecasts. We do not report details of their responses here.

[6] We use generalized eta squared (*ges*) to measure effect size (Olejnik & Algina, 2003).

[7] Aleatory uncertainty refers to uncertainty arising from fundamentally random factors in the environment and is contrasted with epistemic uncertainty that can, in principle, be eliminated by the provision of additional information (Tannenbaum et al., 2017).

[8] There was a significant effect of forecast type on MAE, the alternative measure of overall error, F (1, 85) = 6.15, p = .015, *ges* = 0.0033. See endnote 1.

[9] ME derived from the average of the two point forecasts is the same as that obtained by taking the average of the ME scores calculated for each forecast separately. But, however they were calculated, ME scores were found to be higher with more volatile series, F (1, 44) = 5.82, p = .02, *ges* = 0.0204.

[10] Hypotheses and findings from the three experiments are summarized in Table B1.

REFERENCES

Armstrong, J. S. (2001). *Principles of forecasting: A handbook for researchers and practitioners*. Kluwer Academic.

Benson, P. G., & Önkal, D. (1992). The effects of feedback and training on the performance of probability forecasters. *International Journal of Forecasting*, *8*(4), 559–573.

Block, R. A., & Harper, D. R. (1991). Overconfidence in estimation: Testing the anchoring and adjustment hypothesis. *Organizational Behavior and Human Decision Processes*, *49*, 188–207.

Bruine de Bruin, W., Manski, C. F., Topa, G., & van der Klaauw, W. (2011). Measuring consumer uncertainty about future inflation. *Journal of Applied Econometrics*, *26*(3), 454–478.

Bruine de Bruin, W., van der Klaauw, W., & Topa, G. (2011). Expectations of inflation: The biasing effect of thoughts about specific prices. *Journal of Economic Psychology*, *32*(5), 834–845.

Bryan, M. F., & Venkatu, G. (2001a, Oct 15). *The demographics of inflation opinion surveys*. Federal Reserve Bank of Cleveland, Economic Commentary.

Bryan, M. F., & Venkatu, G. (2001b, Nov 1). *The curiously different inflation perspectives of men and women*. Federal Reserve Bank of Cleveland, Economic Commentary.

Budescu, D. V., & Yu, H. T. (2007). Aggregation of opinions based on correlated cues and advisors. *Journal of Behavioral Decision Making*, *20*(2), 153–177.

Engelberg, J., Manski, C. F., & Williams, J. (2009). Comparing the point predictions and subjective probability distributions of professional forecasters. *Journal of Business and Economic Statistics*, *27*(1), 30–41.

Fischhoff, B. (1994). What forecasts (seem to) mean. *International Journal of Forecasting*, *10*(3), 387–403.

Georganas, S., Healy, P. J., & Li, N. (2014). Frequency bias in consumers' perceptions of inflation: An experimental study. *European Economic Review*, *67*(C), 144–158.

Hansson, P., Juslin, P., & Winman, A. (2008). The role of short-term memory capacity and task experience for overconfidence in judgment under uncertainty. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *34*(5), 1027–1042.

Harvey, N. (1995). Why are judgments less consistent in less predictable task situations? *Organizational Behavior and Human Decision Processes*, *63*(3), 247–263.

Harvey, N., Ewart, T., & West, R. (1997). Effects of data noise on statistical judgment. *Thinking and Reasoning*, *3*(2), 111–132.

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science*, *20*(2), 231–237.

Herzog, S. M., & Hertwig, R. (2014). Harnessing the wisdom of the inner crowd. *Trends in Cognitive Sciences*, *18*(10), 504–506.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. William Collins.

Lawrence, M., Goodwin, P., O'Connor, M., & Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, *22*(3), 493–518.

Lawrence, M., & O'Connor, M. (1992). Exploring judgmental forecasting. *International Journal of Forecasting*, *8*(1), 15–26.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of subjective probabilities: The state of the art up to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge University Press.

Niu, X., & Harvey, N. (2021). Context effects in inflation surveys: The influence of additional information and prior questions. *International Journal of Forecasting*. Published online September 3, 2021.

Niu, X., & Harvey, N. (2022). Outcome feedback reduces over-forecasting of inflation and overconfidence in forecasts. *Judgment and Decision Making*, *17*(1), 124–163.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., & Rakow, T. (2006). *Uncertain judgments: Eliciting experts' probabilities*. Wiley.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, *8*(4), 434–447.

Ramos, M. H., Van Andel, S. J., & Pappenberger, F. (2013). Do probabilistic forecasts lead to better decisions? *Hydrology and Earth System Sciences*, *17*(6), 2219–2232.

Reips, U.-D. (2002). Context effects in web-surveys. In B. Batinic, U.-D. Reips, & N. Bosnjac (Eds.), *Online social sciences*. Hogrefe and Huber.

Roulston, M. S., & Kaplan, T. R. (2009). A laboratory-based study of understanding of uncertainty in 5-day site-specific temperature forecasts. Meteorological Applications. *A Journal of Forecasting, Practical Applications, Training Techniques and Modelling*, *16*(2), 237–244.

Russo, J. E., & Schoemaker, P. J. (1992). Managing overconfidence. *Sloan Management Review*, *33*, 7–17.

Smyth, J. D., Dillman, D. A., & Christian, L. M. (2009). Context effects in internet surveys: New issues and evidence. In A. D. Joynson, K. Y. A. McKenna, T. Postmes, & U.-D. Reips (Eds.), *Oxford handbook of internet psychology*. Oxford University Press.

Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 299–314.

Tannenbaum, D., Fox, C. R., & Ülküman, G. (2017). Judgment extremity and accuracy under epistemic vs, aleatory uncertainty. *Management Science*, *63*(2), 497–518.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. Cambridge University Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Van Dolder, D., & van den Assem, M. J. (2018). The wisdom of the inner crowd in three large natural experiments. *Nature Human Behavior*, *2*(1), 21–26.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*(7), 645–647.

Yaniv, I., & Foster, D. (1995). Graininess of judgment under uncertainty. *Journal of Experimental Psychology: General*, *124*(4), 424–432.

Yates, J. F. (1990). *Judgment and decision making*. Prentice Hall.

Yates, J. F. (1994). Subjective probability accuracy analysis. In G. Wright, & P. Ayton (Eds.), *Subjective probability* (pp. 381–410). Wiley.

Yates, J. F., Price, P. C., Lee, J.-W., & Ramirez, J. (1996). Good probabilistic forecasters: The 'consumer's' perspective. *International Journal of Forecasting*, *12*(1), 41–56.

How to cite this article: Niu, X., & Harvey, N. (2022). Point, interval, and density forecasts: Differences in bias, judgment noise, and overall accuracy. *Futures & Foresight Science*, 1–18. https://doi.org/10.1002/ffo2.124

APPENDIX A

Experiment 1: Uncertainty estimation in interval and density forecasting

It is well established that interval forecasts tend to be much too narrow, thus implying that forecasters are highly overconfident. For example, Hansson et al. (2008) point out: "If people, for example, produce intuitive 90% confidence intervals for unknown quantities, the percentage of intervals that include the true value is often closer to 40% or 50% than to the normatively expected 90% (see Block & Harper, 1991; Lichtenstein et al., 1982; Russo & Schoemaker, 1992; Soll & Klayman, 2004)."

In the interval condition of our experiment in which people were asked to set a 90% interval for each of 10 inflation series, the outcome should be within the set interval in nine of those 10 series. In fact, however, outcomes were within the set interval in an average of only 5.265 series out of 10 (52.65%). A one-sample t test showed that this was significantly below 90%, t (41) = 9.67; $p < .001$. This result, therefore, replicates the overconfidence found in the previous work cited above.

Hansson et al.'s (2008) experiments showed that if, instead of setting an interval for a given probability, people were asked to provide a probability for a given interval, overconfidence all but vanished. In our density condition, we required people to provide a probability equivalent (i.e., a number of pounds sterling out of a maximum of £100) for each of a number of intervals (i.e., bins). As this corresponds closely to the task examined by Hansson et al. (2008), we expected little, if any, overconfidence to be present.

For each series, we first calculated the optimal forecast; in the seven untrended series, this corresponded to the series mean; in the three series with shallow trends, it corresponded to an extrapolation of the trend to the period to be forecasted. After excluding one series (Kiribati) because points were not normally distributed around the mean or trend line, we calculated the 90% prediction interval for each series. The judged probability of the outcome being within this interval should be 90%. In other words, the model fitted to each participant's data on each trial should show that £90 out of the available £100 was allocated to that interval. In fact, we found that, on average, a total sum of £82.92 out of the £100 was allocated to the 90% interval. A one-sample t test showed that this mean value was different from £90, t (30) = 5.90; $p < .001$. Thus, some overconfidence was still present in the density forecasting group.

A two-sample t test showed that the mean probability equivalent assigned to the 90% interval in the density forecasting condition (i.e., 82.92%) was significantly different from the mean probability of outcomes appearing within the judged 90% interval in the interval condition (i.e., 52.65%), t (48.69) = 7.48; $p < .001$. Thus, although interval and density forecasts do not differ in terms of the accuracy with which they provide an estimate of the expected value of the point to be forecast, density forecasts provide a better estimate of the uncertainty associated with the prediction.

APPENDIX B

Table B1

TABLE B1 A summary table of the hypotheses and results in the three experiments

Hypotheses	Experiment	Results
H_1: Point forecast = Mid-point of the interval forecast = Mean of the density forecast	Experiment 1, Experiment 2	Supported
H_2: ME > 0	Experiment 1, Experiment 2	Supported
H_3: $ME_{Point} = ME_{Interval}$ ($=ME_{Density}$)	Experiment 1, Experiment 2	Supported
H_4: $VE_{Point} > VE_{Interval}$ ($>VE_{Density}$)	Experiment 1, Experiment 2	Partially supported: $VE_{Point} > VE_{Interval} = VE_{Density}$
H_5: $RMSE_{Point} > RMSE_{Interval} > RMSE_{Density}$	Experiment 1	Partially supported: $RMSE_{Point} > RMSE_{Interval} = RMSE_{Density}$
H_6: $\lvert VE_{Interval\ second} - VE_{Interval\ first}\rvert > \lvert VE_{Point\ second} - VE_{Point\ first}\rvert$	Experiment 2	Not supported
H_7: $Confidence_{Point\ Forecasts} < Confidence_{Interval\ Forecasts}$	Experiment 2	Supported
H_8: VE of average of two point forecasts < Average VE of two point forecasts	Experiment 3	Supported
H_9: RMSE of average of two point forecasts < Average RMSE of two point forecasts	Experiment 3	Supported
H_{10}: VE of average of two point forecasts = VE of the mid-point of interval forecast bounds	Experiment 3	Supported
H_{11}: RMSE of average of two point forecasts = RMSE of the mid-point of interval forecast bounds	Experiment 3	Not supported
H_{12}: Average VE of two point forecasts > VE of the mid-point of interval forecast bounds	Experiment 3	Supported
H_{13}: Average RMSE of two point forecasts > RMSE of the mid-point of interval forecast bounds	Experiment 3	Supported
H_{14}: The effect identified in H_8 and H_9 will be greater with noisier inflation series	Experiment 3	Supported
H_{15}: The effect identified in H_{12} and H_{13} will be greater with noisier inflation series	Experiment 3	Supported only for H_{13}