Judgment and Decision Making, Vol. 17, No. 1, January 2022, pp. 124–163

Outcome feedback reduces over-forecasting of inflation and overconfidence in forecasts

Xiaoxiao Niu* Nigel Harvey†

Abstract

Survey respondents over-forecast inflation: they expect it to be higher than it turns out to be. Furthermore, people are generally overconfident in their forecasts. In two experiments, we show that providing outcome feedback that informs people of the actual level of the inflation that they have forecast reduces both over-forecasting and overconfidence in forecasts. These improvements were preserved even after feedback had been withdrawn, a finding that indicates that they were not produced because feedback had a temporary incentive effect but because it had a more permanent learning effect. However, providing forecasters with more outcome feedback did not have a greater effect. Feedback appears to provide people with information about biases in their judgments and, once they have received that information, no additional advantage is obtained by giving it to them again. Reducing over-forecasting also had no clear effect on overall error. This was because providing outcome feedback after every judgment also affected the noise or random error in forecasts, increasing it by a sufficient amount to cancel out the benefits provided by the reduction in over-forecasting.

Keywords: over-forecasting, overconfidence, feedback, judgment noise, inflation surveys, inflation expectations

1 Introduction

Human performance of high-level cognitive tasks such as the forecasting (e.g., Lawrence et al., 2006), monitoring (e.g., Brown and Steyvers, 2009; Fiedler et al., 2016), and control (e.g., Frensch and Funke, 1995; Kerstholt, 1996; Osman, 2010) of complex systems is typically sub-optimal. All these tasks involve judgment and many methods of enhancing its quality have been studied. They include application of financial (Camerer and Hogarth, 1999) and social (Lerner and Tetlock, 1999) incentives, training using guidance (e.g., Chang

*Department of Experimental Psychology, University College London.
†Department of Experimental Psychology, University College London. Email: n.harvey@ucl.ac.uk.

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

et al., 2016; Mellers et al., 2014; Nisbett, 1993) and feedback (e.g., Hammond et al., 1973; Harvey, 2011; Sterman, 1989), provision of advice (Bonaccio and Dalal, 2006) and technical support (Edwards and Fasolo, 2001), use of groups or teams rather than individuals (e.g., Sniezek, 1990), and recoding of information to render it easier to process (e.g., Sedlmeier, 2000). The effectiveness of these approaches varies considerably: for example, depending on the task, provision of feedback or incentives may improve, have no effect on, or impair performance.

Forecasting is the most ubiquitous of these high-level tasks. There is hardly an area of modern life from which it is absent: economics, finance, meteorology, traffic, energy, demographics, business, tourism, sports, elections, geopolitics, and epidemiology are just a few of the domains where it is critical for planning (Petropoulos et al., 2022). Nowadays, much forecasting depends on a combination of algorithmic and judgmental processing. However, unaided judgment is still used: for example, a recent survey of demand (sales) forecasting in businesses (Fildes & Petropoulos, 2015) showed that this approach is still employed by 16% of companies. Here we investigate whether providing people with outcome feedback (OFB) after they have made their forecasts increases their accuracy and makes their confidence in their performance more realistic.

Unaided judgmental forecasting is required in the surveys of inflation expectations that are carried out as part of the process used by central banks to forecast inflation and other important economic variables (e.g., unemployment levels, interest rates, etc.). Given its importance, we framed our studies within this forecasting domain. However, we would expect our findings to generalize to other types of judgmental forecast (e.g., sales forecasts).

In what follows, we first explain why central banks consider it important to survey laypeople's inflation expectations. We go on to review studies of the effects of OFB on a) judgmental forecasting and b) confidence in judgments. We then present the results of two experiments. The first of these compares forecasts and confidence in forecasts under a condition in which OFB is provided after every forecast with one in which it is never provided. The second one examines the effects of varying the amount of OFB (after every judgment, after every other judgment, etc.).

1.1 Inflation expectations

Central banks use surveys, such as the Michigan Survey of Consumers (MSC), to obtain information about laypeople's inflation expectations because these expectations are assumed to influence future inflation levels. For example, the more that people expect inflation to increase, the more they will bring forward their planned purchase of durable goods. This, in turn, will push up the price of those goods by increasing demand for them and, hence, will increase inflation.

Lay inflation expectations are typically much too high (Bruine de Bruin et al., 2011; Bryan and Venkatu, 2001a, b; Georganas et al., 2014). One reason for this is that people generally expect the inflation rate (as distict from prices) to go up, even when it does not do

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

so. Niu and Harvey (2021) showed that when separate groups of people estimated current inflation and inflation for the next year, their mean estimates for the two years were not significantly different. However, when the same group of people made estimates for both years, their mean estimate for the next year was significantly higher than their mean estimate for the current year.

Central banks also use surveys, such as the Survey of Professional Forecasters (SPF), to gain information about professional forecasters' inflation expectations. These also show biases though they are typically smaller than those found among lay respondents (Lei et al., 2015). The nature and size of these biases change over time (Kim and Kim, 2009; Schuh, 2001), depending on the level of inflation and on whether it is trending upwards to downwards (Lei et al., 2015). However, Ang et al., (2007, p1178) conclude that "the SPF survey mostly over-predicts inflation".

It is reasonable to assume that estimates of inflation would be of more use to central banks if they were more accurate. Here we examine whether they can be improved by providing people with training using OFB. In other words, after each inflation forecast, people were told what the actual value of inflation turned out to be. We were able to provide this feedback by using real historical inflation data for 20 different countries. (Countries were anonymised to ensure that forecasters used only the data presented to them.)

Before presenting our experiments, we review previous work on the effects of OFB on two aspects of judgmental forecasting. First, feedback may affect the accuracy of forecasts. Second, it may affect the confidence that people have in their forecasts. Both of these are important for planning purposes. For example, if I expect that demand for a product will increase, I may develop additional manufacturing capability to meet that demand but I would be less likely to make major changes if I were less confident in my forecast. Similarly, if a central bank survey shows that people expect inflation to increase, the bank would take this into account when making their inflation forecasts but the weight that they put on this information would depend on people's confidence in their inflation expectations. In the next sections, we consider work on the effects of OFB on the three types of forecasting (memory-based, cue-based, time series) first with respect to accuracy and then with respect to confidence.[1]

1.2 Effects of outcome feedback on forecast accuracy

Judgmental forecasting relies on information held in memory, on information from a single point in time about values of a set of variables other than the one to be forecast, or on information about past values of the variable to be forecast. We term these memory-based forecasting, cue-based forecasting, and time series forecasting, respectively. Though

[1]We restrict our discussion to point forecasts (e.g., Inflation will be 2.25% next year) and event forecasts (e.g., Inflation will increase next year). We do not cover prediction intervals (e.g., There is 90% likelihood that inflation will be between 2.00% and 2.80%) or probability forecasting (e.g., There is a 90% likelihood that inflation will increase next year).

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

particular forecasting tasks carried out by practitioners may draw on more than one of these types of information, they have been studied separately by psychologists. More specifically, different types of heuristic have been used to account for characteristics of performance in each case (Harvey, 2007).

1.2.1 Memory-based forecasting

Studies of memory-based forecasting have focussed on sports forecasting (e.g., Ayton et al., 2011; Pachur and Biele, 2007; Serwe and Frings, 2006), financial stock forecasting (e.g., Andersson and Rakow, 2007; Borges et al., 1999) and geopolitical forecasting (e.g., Chang et al., 2016; Mellers et al., 2014). A major concern has been to identify the nature of the memory-based heuristic used to make forecasts. Examining effects of OFB has not been a major concern. However, Ayton et al.'s (2011) study is relevant to us here. They asked 50 Turkish students to make forecasts for the full-time results of 32 English football matches. Once they had done this, they were given the half-time results for 19 of those matches and then they re-forecast the full-time results. The half-time results provided partial OFB. This information did influence the forecasts that people made but did not increase the accuracy of those forecasts: 62.5% were correct without feedback and 60% were correct with it.

1.2.2 Cue-based forecasting

Research on cue-based forecasting has used the multiple-cue probability learning (MCPL) paradigm (Cooksey, 1996). In MCPL tasks, people estimate the values of a criterion (e.g., future examination grades) for each of a number of instances (e.g., different students) after being given the values of a set of cues for each of those instances (e.g., past examination grades, past marks from continuous assessment, number of absences from teaching classes).

Many studies have shown that, when more than two or three cues are given or when the relation between cues and criterion is non-linear, OFB fails to produce learning (Deane et al., 1972; Goldberg, 1968; Hammond, 1971; Hammond and Boyle, 1971; Hammond and Summers, 1972). In more complex tasks, it can impair performance (Hammond et al., 1973; Holzworth and Doherty, 1976; Schmitt et al., 1976; Schmitt et al., 1977). Reviewers (e.g., Klayman, 1988; Karelaia and Hogarth, 2008) confirm this impression.

There has been some debate about the reasons for the ineffectiveness of OFB in MCPL tasks. In these tasks, a random error term is added to (or is assumed to be present in) the expression that generates the criterion from the cue values. Brehmer (1980, p233) claimed that "people simply do not have the cognitive schemata needed for efficient performance in probabilistic tasks". In contrast, Todd and Hammond (1965) suggested that OFB fails because it does not specify what people need to do in order to improve their performance: finding that a forecast for an examination was too high does not tell the forecaster how to change the weights on the different cues. However, in advice tasks where cues are suggestions for the values of the criterion made by different advisors, OFB does provide

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

information about how to re-weight the cues. The finding that OFB does permit more rapid learning in such tasks than in MCPL tasks (Fischer and Harvey, 1999; Harries and Harvey, 2000) favours Todd and Hammond's (1965) approach over that of Brehmer (1980).

1.2.3 Time series forecasting

There is a problem with studying OFB in time series forecasting. It is that, for a given time series, recurrent forecasting of a particular series means that the actual outcome that has just been forecast must be revealed so that it can be used when making the next forecast. Thus, OFB is always present. As a result, OFB conditions have been used as a baseline control against which the effectiveness of more elaborate types of feedback, such as performance feedback[2], has been assessed (e.g., Goodwin and Fildes, 1999; Remus et al., 1996). In considering the effectiveness of these different forms of feedback, Lawrence et al. (2006, pp507–8) conclude that "studies have tended to show that outcome feedback is the least effective form" and, following Klayman (1988), they suggest that this "is probably because the most recent outcome contains noise and hence it is difficult for the forecaster to distinguish the error arising from a systematic deficiency in their judgement from the error caused by random factors". In other words, OFB is relatively ineffective because forecasters have difficulty filtering out the contribution that system noise makes to the values of the realized outcomes of the variable being forecast.

The question that remains to be answered is whether OFB is effective to any degree. Does it facilitate forecasting, impair forecasting, or have no effect whatsoever? As far as we can ascertain, no studies have addressed this issue. Yet it has clear relevance to forecasting practice. For example, after sales forecasts have been made and used for planning purposes, it is often the case that they are not retained. Is this a sensible strategy or would improvements in forecast accuracy produced by OFB be worth the additional costs of retaining forecasts and comparing them with outcomes? To determine whether OFB is effective, the standard paradigm of requiring recursive forecasts from the same data series is of no use for the reasons that we outlined above. Instead, a single forecast (or a set of simultaneous forecasts) must be made from each of a number of data series by participants who either receive OFB or do not do so. It is this approach that we take in the experiments reported below.

1.3 Effects of outcome feedback on confidence in forecasts

Confidence in forecasts is usually assessed by asking people to estimate the probability that their forecast will be correct or that it will fall within certain pre-specified bounds. The probabilities they give can then be compared with objective frequencies obtained from a sample of their judgments. For each forecast, f, (e.g., there is a 60% chance of my sales forecast being within 20% of the actual sales), we set an outcome index, d, at 1.00 when

[2]For example, forecasters could be informed of their root mean squared prediction error.

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

the event occurs (the forecast is within 20% of the outcome) and at 0.00 when the event does not occur. The probability score, PS, is then calculated from the forecast (expressed as a probability rather than a percentage) as follows: $PS = (f - d)^2$. The probability score is also known as the Brier score (Brier, 1950) and gives a measure of the quality of these confidence judgments with lower mean probability scores indicating better judgments. Someone who judges that all forecasts as likely to be correct as not and therefore always provides a probability judgment of 0.5 (a uniform judge) obtains a mean probability score of 0.25. Overconfidence is measured by the mean value of $f - d$.

Brier scores, together with bias scores, give us a way of assessing the overall accuracy of probability judgments. Brier scores are often referred to a measure of 'calibration-in-the-large'. With a sufficient data points per forecaster, they can be usefully decomposed into components. The commonly used Murphy decomposition divides the Brier score into incidence (a measure of the relative frequency of the target event), discrimination or resolution (a measure of ability to distinguish between target and non-target events), and calibration (a measure of ability to match labels for subjective probability to values of objective frequencies). The covariance decomposition (Yates, 1982; Yates & Curley, 1985) is useful for providing insights into the psychological processes underlying probability estimation. It divides the mean probability score into incidence, bias, separation (mean forecast when target event is present minus mean forecast when it is absent), and scatter (judgment noise).

1.3.1 Overconfidence

Much research indicates that a) people are typically overconfident in their decisions, and b) overconfidence is greater with harder decisions, with under-confidence present with very easy ones (the hard-easy effect). However, much of the original research supporting these conclusions derived from studies that required people to give a probability (50–100%) that each of their answers to a set of two-alternative general knowledge questions was correct (e.g., Lichtenstein et al., 1982). Later work (e.g., Gigerenzer et al., 1991; Juslin, 1994) implies that these findings arose because experimenters had included too many misleading questions (i.e., those with counterintuitive answers) in their experiments. When questions were made more representative of the natural ecology by randomly selecting them from a population of all possible questions of a particular type, overconfidence greatly diminished. However, overconfidence and the hard-easy effect have been found in many other tasks where the ecological critique is harder to sustain (Harvey, 1994). These include perceptual tasks (e.g., Baranski & Petrusic, 1994) and motor tasks (e.g., Cohen et al., 1956; Cohen & Dearnley, 1962). However, in these reports, the effects do not appear to be as large as those reported in the original studies of confidence in answers to general knowledge questions.

There has been some debate about whether people are overconfident in their forecasts. Fischhoff and MacGregor (1982) found that confidence in forecasts had similar characteristics to confidence in answers to general knowledge questions (i.e., general overconfidence

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

together with a hard-easy effect), though there were fewer 100% certain responses in the former case. However, Wright (1982) found that people were less overconfident in their forecasts about future events than in their answers to equally difficult questions about whether similar events had already occurred. Also, Ronis and Yates (1987) found that people were less overconfident in their forecasts of basketball games than in their answers to general knowledge questions. In summary, it appears that forecasters are still biased towards overconfidence but not to the same degree as those answering general knowledge questions. This difference is likely to reflect the validity of the ecological critique of studies using general knowledge questions (Gigerenzer et al., 1991; Juslin, 1994).

1.3.2 Effect of outcome feedback on overconfidence in answers to related and unrelated questions

A few studies have examined whether OFB reduces overconfidence or otherwise improves the quality of confidence judgments. For example, over four weekly sessions, Benson and Önkal (1992) asked people to forecast which team would win in each of 55 major college football games and then to express their confidence in their forecast as a probability between 0.5 and 1.0. In one condition, people received OFB at the start of the last three sessions. This listed their predictions and probability assessments from the previous week, along with the actual scores in the 55 games. Results showed that the mean probability score, calibration, and scatter deteriorated over the four weeks and that there was no change in resolution or slope. However, bias in confidence judgments was reduced. This is certainly suggestive of an effect of OFB. Unfortunately, there was no control group in which OFB was not given: the effects reported could have occurred as a function of experience over time in the absence of OFB.[3]

Subbotin (1996) asked people to go through a list of 100 pairs of European capitals or 100 pairs of European countries and, for each one, a) decide which was larger and then b) estimate the probability (50%-100%) that their answer was correct. When decisions were easy, people were under-confident and simple OFB reduced this bias. However, when decisions were difficult, people were overconfident and simple OFB had no effect. Furthermore, when Subbotin (1996) repeated his experiment using unrelated questions that did not all refer to the same variable (city or country size), OFB had no effect irrespective of question difficulty.

In Winman and Juslin's (1993) experiment, people in one condition answered a series of related questions about which of two causes of death was the more common among Swedes and those in another condition decided which was the heavier of two weights. After each answer, people in both groups estimated the likelihood it was correct (50–100%). Feedback

[3]Various studies, including Benson and Önkal (1992), have examined effects of other types of feedback in which people were told of their accuracy, their calibration scores, their resolution scores, and other highly processed performance measures (e.g., Adams & Adams, 1958; Lichtenstein & Fischhoff, 1980; Sharp et al., 1988; Stone & Opel, 2000). Obtained effects have generally been modest (Keren, 1991).

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

was given after each of the 40 trials in the middle four of six trial blocks. (Participants were told whether their decision had been right or wrong and, so, strictly speaking, they were given performance feedback rather than OFB.) There was also a control condition in which people did the causes of death task without any feedback. An initial bias towards overconfidence in the causes of death task disappeared by the fourth block when feedback was given but remained throughout the experiment when it was not supplied. However, feedback had no effect on the under-confidence bias observed in the weight discrimination task.

Keren (1988) required people to carry out two tasks with or without OFB. In one, they answered unrelated general knowledge questions and, after each response, gave a confidence rating (50–100%). In the other, they decided whether a gap in a perceptual stimulus was on the left or right and gave the same type of confidence rating. This perceptual task was hard (small gaps with 67% correct and slight under-confidence) or easy (large gaps with 79% correct and high under-confidence). The difficulty of the general knowledge task was midway between that of the perceptual tasks (71% correct with overconfidence). OFB had no effect on under-confidence in the perceptual tasks, whatever their difficulty, or on overconfidence in the general knowledge task.[4]

In their review of this issue, Russo and Schoemaker (1992, p. 11) state that: "We believe that timely feedback and accountability can gradually reduce the bias toward overconfidence in almost all professions. *Being 'well-calibrated' is a teachable, learnable skill*" (italics are theirs). However, in commenting on this statement, McClelland and Bolger (1994, p. 476) argue that "the evidence that outcome feedback alone is effective in reducing miscalibration is not encouraging". Nevertheless, this evidence has some structure. OFB has no effect on biased confidence in answers to unrelated general knowledge questions (Keren, 1988; Subbotin, 1996) or on biased confidence in perceptual decisions (Keren, 1988; Winman & Juslin, 1993). However, OFB has been found to decrease biased confidence in answers to *related* questions (Benson & Önkal, 1992; Subbotin, 1996; Winman & Juslin, 1993).

One report does not fit this neat picture. Zakay (1992) examined groups of people who answered general knowledge questions and who, after each answer, estimated the probability (0.5–1.0) that it was correct. Half carried out the experiment as a pencil-and-paper test and, for the other half, the experiment was run on computers. Within each of these groups, half the people received OFB after each answer and the other half did not. Feedback reduced overconfidence in the computerized version of the task but not in the paper-and-pencil version. The result in the former case is an outlier in that all other reports of experiments with unrelated items failed to show an effect of feedback on overconfidence. Moreover, the interaction was barely significant at the 5% level, and each group had only 10 participants.

[4]In addition, Önkal and Muradoğlu (1995) and Fischer (1982) found no effects of OFB in full-range probabilistic forecasting tasks. In such tasks, people simply assess the probability (0–100%) that an event will happen; no separate confidence judgment is made. Keren (1991) recommends avoiding use of these tasks in the study of confidence because they suggest to participants that they are to consider aleatory uncertainty inherent in the system being assessed rather than the epistemic uncertainty present in their own minds.

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

1.4 Identifying the nature of outcome feedback effects

OFB may have two main effects (Annett, 1969). First, it may incentivise people to perform better: they may put more effort into tasks when they know they will find out how well they have performed. This effect disappears when feedback is removed. Second, feedback may produce learning that improves people's forecasting ability: this effect is maintained when feedback is removed. To distinguish between these effects, two groups of participants must be used in experiments. One group should receive OFB in a first session but not in a second session; another group should receive feedback in neither session. A difference between the two groups in the first session shows a beneficial effect of feedback but does not identify the nature of this effect. Disappearance of the beneficial effect of feedback in the second session indicates that it was due to incentivisation. Evidence that it is fully maintained implies that it was due to learning. Some partial maintenance of the effect would suggest that it arose because both incentivisation and learning had a role in producing it.

Incentivisation increases speed or accuracy in those tasks in which greater mental effort is effective in improving performance (Camerer & Hogarth, 1999; Lerner & Tetlock, 1999). The possibility that this explains OFB effects (e.g., by increasing the attention paid to critical aspects of the task) has been recognised (e.g., Adams & Adams, 1958; Zakay, 1992). However, only Winman and Juslin (1993) used an experimental design of the sort described above to determine whether incentivisation or learning produced the OFB effects that they obtained. They found that these effects were fully maintained after feedback was removed, thereby indicating that they were produced by learning.

1.5 Rationale and hypotheses

We report two experiments designed to determine whether OFB reduces laypeople's overestimation of inflation and whether it reduces their overconfidence in their inflation estimates. The experiments use the design outlined above to determine whether any obtained effects arise from training-induced learning rather than from incentivising effects of OFB.

On the basis of the previous work reviewed above, our first hypothesis (H_1) was that lay estimates of inflation will be too high. We are interested in whether OFB reduces this overestimation. We have seen that studies have shown that there is no evidence that OFB improves memory-based forecasts. There is some evidence that it benefits cue-based forecasts but only when few cues are combined in a simple way. No studies of effects of OFB on time series forecasting have been reported. However, as this type of forecasting is closer to cue-based forecasting than to memory-based forecasting (because some external information is provided), we tentatively hypothesise that OFB will improve time series forecasting (H_2).

We have seen that people tend to show overconfidence in their responses to related items (though this overconfidence is not as great as when they respond to unrelated items). Our participants made forecasts for each member of a sequence of time series showing inflation

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

rates in different countries. As these were related items, we expected a modest degree of overconfidence in these forecasts (H_3). Furthermore, we saw above that evidence indicates that OFB reduces overconfidence in responses to related items. As our task involved responses to related items, we expected it to have that effect here. Specifically, we expected it to lower confidence and, as a result, reduce bias and improve calibration (H_4). Finally, because of the hard-easy effect, we considered that overconfidence would be greater for series that were more difficult to forecast (H_5).

2 Experiment 1

Laypeople were presented with a tabular record of ten years (2009–2018) of inflation figures (Consumer Price Index, CPI)[5] and required to make a one-step ahead forecast for 2019 and to express their confidence (0–100%) that this forecast was within 20% of the actual level of inflation. They carried out this task for a sequence of 20 anonymised countries. The Feedback group received OFB for the first set of ten countries (session 1) but not for the second set of ten (session 2). The No-feedback group received no feedback in either session. (We use the term 'session' rather than 'trial block' because the two sets of trials were separated by a new instruction page that informed participants that they would not receive feedback information and that they would now be given a bonus payment related to their performance.)

2.1 Method

2.1.1 Participants

One hundred and two participants (34 females, 68 males) with a mean age of 27 years (SD = 9 years). They were recruited via the web platform Prolific.com between 24 and 27 October 2020. They were paid £ 1.10 each for their participation. Additionally, in the second session, they were given a bonus of £0.10 whenever their inflation estimate for a country was within 20% of the correct value. The study was approved by UCL Department of Experimental Psychology Ethics Committee.

2.1.2 Design

This was a mixed design experiment. The between-participants factor was participant group (Feedback group versus No-feedback group) and the within-participant factors were Session (first versus second session) and Trial (1–10). OFB was provided only to the Feedback group in the first session. Within each session, participants made judgments for ten countries that were presented in a different random order for each participant.

[5]Surveys such as the Federal Reserve Bank of Philadelphia's SPF provide respondents with values of inflation prior to the year for which their forecast is required.

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

2.1.3 Stimulus materials

Eleven years (2009–2019) of historical data were extracted from the Word Bank dataset of annual consumer price inflation (CPI) for 20 countries (Armenia, Australia, Belgium, Brazil, China, Colombia, Denmark, EI Salvador, Fiji, Finland, France, Greece, India, Ireland, Japan, South Korea, Peru, Thailand, the United Kingdom and the United States). On each trial, the first ten of these figures (2009–2018) were shown in a table for one of the countries. Figure 1 shows an example of one of these tables. For each participant, ten out of 20 countries were randomly allocated to the first session and randomly ordered within that session. The remaining 10 countries were randomly ordered to form the second session.

First set of ten countries: Country 1

Task instructions

Please provide your estimation for Consumer Price Inflation rate (2019) in this table by typing in the one blank cell, which should be computed at the annual-average level.

Please give your prediction using two figures after a decimal point: for example, 20.47, 14.66 or 0.00.

Consumer price inflation (annual %)

	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Country O	–1.35	–0.72	–0.27	–0.05	0.35	2.76	0.79	–0.12	0.47	0.98	

FIGURE 1: A table showing annual inflation figures between 2009 and 2018 and an empty cell for participants to enter their inflation estimate for 2019.

Participants made inflation estimates by completing the 2019 cell at the right-hand end of the table. After doing this, they were asked to move a slider to provide their judgment of the likelihood (0–100%) that their inflation estimate was within 20% of the correct value. In the first session, participants in the experimental group then moved on to a feedback screen. As an example, consider a trial in which the participant had estimated inflation for 2019 to be 1.11% when it was actually 0.48%. Feedback was given in two lines of text. In this example, the upper one would have read "The actual consumer price inflation (2019) for this country is 0.48%" and the lower one would have read "You estimated consumer price inflation (2019) for this country to be 1.11%". In the second session, participants in the experimental group did not receive feedback but moved directly on to the next trial. Participants in the No-feedback group did not receive feedback in either session.

2.1.4 Procedure

After reading the information sheet and completing the consent form, participants read the task instructions. These instructions were also provided on each trial (Figure 1). Before starting the task, participants were provided with a simple definition of consumer price inflation, together with an example.

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

In the first session, participants in the Feedback group but not those in the No-feedback group received OFB in the format described above. In the second session, participants in neither group received OFB. At the end of the experiment, participants provided demographical details, including their age, gender, highest level of education qualification attained, economic related knowledge and work experience.

2.2 Results

Fifteen outliers whose judgments were more than three standard deviations from the mean values of the judgments for the 20 countries were excluded. The sample analysed below therefore comprised 87 participants (59 men, 28 women) with a mean age of 27 years (SD= 8 years). Forty-six of them were in the Feedback condition and 41 of them were in the No-feedback condition.

Trial-by-trial data for inflation judgments and confidence judgments are shown in Table A1 of the Appendix. Means of these inflation judgments and their associated accuracy levels across all 20 countries in the experiment, together with corresponding values produced by two simple algorithmic models, are given in Table A4 of the Appendix.

2.2.1 Forecasting performance

To obtain an overall measure of accuracy for each participant on each trial, we used the absolute error (AE): this is the absolute difference between the relevant value of inflation given by the World Bank for 2019 and the value estimated by the participant. The algebraic difference between the same two values provided our measure of constant error (CE), also known as directional error or bias. Constant error is one of two types of error that contributes to overall error. The other is variable error (VE), also known as scatter, noise, or inconsistency. This variable error adds random noise to the measure of CE that is obtained by taking the algebraic difference between the correct and judged value of inflation. To obtain a measure of the size of VE for each participant on each trial, we proceeded as follows. Separately for each participant in each session, we fitted a trend line with linear and quadratic components to the values of CE. We then extracted the residuals from these regressions and used the absolute value of the residual on each trial as our measure of VE on that trial. Mean values of raw inflation judgments and these three error measures are shown in Table 1 for the two sessions and the two conditions of the experiment.

To address the issue of whether people tend to over-estimate future inflation rates (H_1), we used one-sample t-tests to determine whether the mean CE values in the first and second sessions completed by the Feedback and No-feedback groups were significantly different from zero. In each case, the test was significant: Feedback group, session one ($t(45) = 5.51$, $p < 0.001$); Feedback group, session two ($t(45) = 4.47$, $p < 0.001$); No-feedback group, session one ($t(40) = 5.75$, $p < 0.001$); No feedback group, session two ($t(40) =$

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

TABLE 1: Experiment 1: Means and standard deviations (in parentheses) of inflation judgments, their absolute errors, constant errors and variable errors.

	Feedback condition	No-feedback condition	mean
Judgment			
Session 1	2.00(0.42)	2.17(0.40)	2.08(0.41)
Session 2	1.98(0.34)	2.14(0.42)	2.06(0.38)
mean	1.99(0.22)	2.15(0.28)	2.07(1.38)
Absolute error			
Session 1	0.89(0.27)	0.88(0.31)	0.89(0.29)
Session 2	0.87(0.19)	0.88(0.22)	0.87(0.21)
mean	0.88(0.14)	0.88(0.19)	0.88(0.86)
Constant error			
Session 1	0.20(0.25)	0.39(0.43)	0.29(0.36)
Session 2	0.24(0.36)	0.37(0.35)	0.30(0.36)
mean	0.22(0.22)	0.38(0.28)	0.29(1.19)
Variable error			
Session 1	0.81(0.31)	0.69(0.25)	0.75(0.29)
Session 2	0.75(0.22)	0.66(0.21)	0.71(0.22)
mean	0.78(0.16)	0.68(0.12)	0.73(0.68)

6.79, $p < 0.001$). These results confirm those of many previous studies: laypeople tend to overestimate inflation rates.

To examine the effects of OFB on inflation judgments (H_2), we carried out three-way mixed ANOVAs on the raw inflation judgments, AE, CE, and VE with Condition (feedback versus no-feedback) as a between-participant factor and with Session (one versus two) and Trial within Sessions (1–10) as within-participant factors.

The analysis of the inflation judgments showed a main effect of Condition ($F(1, 85) = 9.06$, $p = 0.003$, $ges = 0.0035$)[6]: participants in the no-feedback group estimated inflation to be higher than those in the feedback group. There was also a main effect of Trial ($F(8.10, 688.50) = 2.79$, $p = 0.005$, $ges = 0.0141$). As Table A1 shows, this arose because participants in both groups estimated inflation to be higher at the beginnings and ends of each session.

[6]When Mauchy's test showed a deviation from sphericity, Greenhouse-Geissser corrections were used to adjust degrees of freedom. Generalised eta squared (*ges*) measured effect size (Olejnik and Algina, 2003).

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

Analysis of AE showed no main or interactive effects of any of the three variables (Table 1). Analysis of CE showed only a main effect of Condition (F (1, 85) = 9.03, p = 0.003, *ges* = 0.0046): participants in the no-feedback group overestimated inflation more than those in the feedback group (Table 1). Analysis of VE also showed only a main effect of Condition (F (1,85) = 11.13, p = 0.001, *ges* = 0.0060) but it was in the opposite direction to that observed for CE: participants in the feedback group made noisier judgments than those in the no-feedback group (Table 1).

Taken together, these analyses suggest that feedback had no effect on overall error (AE) because its effects on CE and VE were in opposite directions.

2.2.2 Confidence in forecasts

Table 2 shows mean values of confidence judgments, bias scores, and calibration scores (i.e., 1−Brier score) for both conditions and both sessions. (We transformed the Brier score to produce the calibration score so that higher scores indicate better performance.)

TABLE 2: Experiment 1: Means and standard deviations (in parentheses) of confidence judgments, their bias scores, and calibration scores.

	Feedback condition	No-feedback condition	mean
Confidence judgment			
Session 1	48.66(18.91)	56.20(18.91)	52.21(19.18)
Session 2	47.49(20.20)	59.99(19.15)	53.38(20.57)
mean	48.08(19.19)	58.09(18.50)	52.80(23.34)
Bias			
Session 1	0.24(0.24)	0.29(0.22)	0.26(0.23)
Session 2	0.24(0.24)	0.38(0.23)	0.30(0.24)
mean	0.24(0.22)	0.33(0.20)	0.28(0.49)
Calibration			
Session 1	0.70(0.10)	0.66(0.10)	0.68(0.10)
Session 2	0.72(0.12)	0.64(0.14)	0.68(0.13)
mean	0.71(0.10)	0.65(0.11)	0.68(0.24)

To investigate whether people were overconfident in their inflation estimates (H_3), we used one-sample t-tests to determine whether the mean bias scores in the first and second sessions completed by the Feedback and No-feedback groups were significantly different from zero. In each case, the test was significant: Feedback group, session one (t (45) = 6.60, p < 0.001); Feedback group, session two (t (45) = 6.71, p < 0.001); No-feedback group,

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

session one (t (40) = 8.43, p < 0.001); No-feedback group, session two (t (40) = 10.53, p < 0.001). These results are consistent with the hypothesis that laypeople are overconfident in their estimates of inflation rates.

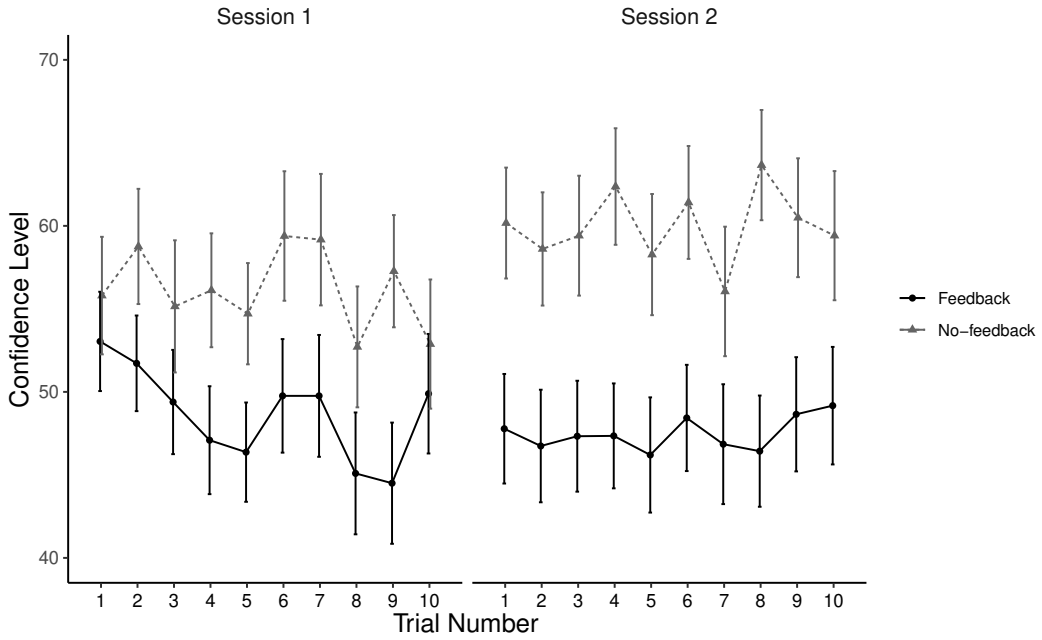

FIGURE 2: Experiment 1: Graph of mean confidence level (with standard error bars) in the two sessions for each group.

To examine the effects of OFB on confidence (H_4), three-way mixed ANOVAs were carried out on the confidence variables using the same factors as those used for the analyses of forecasts. The analysis of confidence level showed an effect of Condition (F (1, 85) = 6.10, p = 0.015, *ges* = 0.0464) and an interaction between Condition and Session (F (1, 85) = 7.85, p = 0.006, *ges* = 0.0030). Further investigation showed that the simple effect of Condition was significant in Session 2 (F (1, 85) = 8.71, p = 0.004, *ges* = 0.0929) but not in Session 1 and that the simple effect of Session was significant for the No-feedback condition (F (1, 40) = 7.41, p = 0.01, *ges* = 0.0101) but not for the Feedback condition. As shown in Figure 2, these effects arose because participants in the No-feedback condition had higher confidence in their forecasts than those in the Feedback condition and because only participants in the No-feedback condition increased their confidence from session one to session two.

Finally, there was an interaction between Trial and Session (F = (7.64, 649.49) = 2.15, p = 0.03, *ges* = 0.0035). This arose because a simple effect of Trial was significant in Session 1 (F (6.89, 592.88) = 2.13, p = 0.04, *ges* = 0.0077) but not in session 2. Confidence dropped over Session 1 but showed no further change over Session 2. These effects are shown in Figure 2.

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

The analysis of Bias showed a main effect of Condition (F (1, 85) = 4.52, p = 0.036, ges = 0.0098): overconfidence was greater in the No-feedback condition than in the Feedback condition. There also an interaction between Condition and Session (F (1,85) = 4.07, p = 0.047, ges = 0.0020). Tests of simple effects showed that the effect of Condition was significant only in session two (F (1, 85) = 7.74, p = 0.007, *ges* =0.0835) and that the effect of Session was significant only in the No-feedback condition (F (1, 40) = 7.79, p = 0.008, *ges* = 0.0370). These effects show that, overall, participants were more overconfident in the No-feedback condition than in the Feedback condition and that this was because only participants in the No-feedback condition showed greater overconfidence in session two than in session one (Table 2).

Analysis of the Calibration scores showed a main effect of Condition (F (1, 85) = 6.50, p = 0.01, ges = 0.0139) and a marginally significant interaction between Condition and Session (F (1, 85) = 3.80, p = 0.054, ges =0.0022). Further analyses showed that the simple effect of Condition was significant only in Session 2 (F (1, 85) = 8.16, p = 0.005, ges = 0.0876) and that the simple effect of Session was not significant for either Condition. These effects are shown in Table 2.

Finally, we investigated whether there was a hard-easy effect (H_5). For the Feedback group, a regression showed that Bias = 0.48 − 0.01·Difficulty (Adj R^2 = 0.97; F (1, 18) = 654.80, p < 0.001); for the No-feedback group, the corresponding regression showed Bias = 0.57 − 0.01·Difficulty (Adj R^2 = 0.95, F (1, 18) = 346.67). It is clear from these analyses that OFB affected the intercept (overall overconfidence) but had no effect whatsoever on the slope (hard-easy effect) as shown in Figure 3.

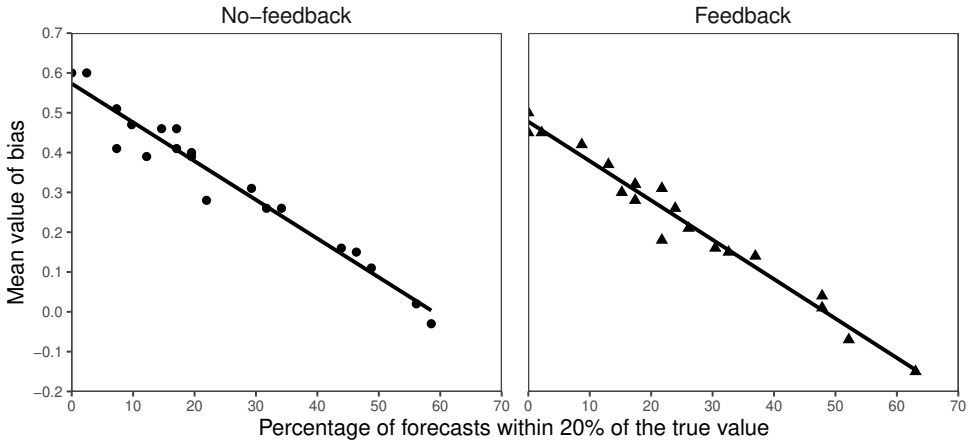

FIGURE 3: Experiment 1: Bias plotted against forecast difficulty in the no-feedback group (left panel) and the feedback group (right panel).

2.3 Discussion

People given the previous ten years of inflation levels for a set of countries overestimated the levels of inflation for the following year in those countries. This finding confirms reports

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

of overestimation of inflation in a number of previous studies (e.g., Bruine de Bruin et al., 2011). However, in those studies, people were typically required to estimate future inflation without information about previous levels of inflation. We have shown that, even with that information, overestimation still occurs. However, provision of OFB over a set of 10 inflation judgments almost halved the degree of the overestimation. Furthermore, this effect continued once feedback had been removed. We conclude that it was produced via learning rather than incentivisation.

Despite this beneficial effect of OFB, overall error did not decline. This was because feedback had a detrimental effect in addition to its beneficial influence on bias: it increased the scatter or random error in judgments. This source of error in judgments has been recognised for some time (Thurstone, 1926) and has been examined thoroughly by judgment analysts working on multiple-cue probability learning (e.g., Brehmer, 1978). It explains why people make different judgments when presented with the same information on different occasions. However, we did not expect feedback to increase variable error. What could have produced this effect?

When people make a sequence of forecasts from a volatile series of independent points, their forecasts should lie on the mean value of, or on the trend line through, those points. Instead, they scatter their sequence forecasts around the mean or trend line. People making forecasts from time series tend to add noise to those forecasts; the noise they add is higher when the series noise is higher but still somewhat less than the series noise (Harvey, 1995; Harvey et al., 1997). It appears that forecasters aim to simulate the noise (as well as the pattern) in the series but do not fully succeed in doing so. It appears that providing people with feedback gives them the impression that the volatility in the series from which they had forecast was greater than they had assumed when making their forecast. As a result, they attempt to simulate the noise in the series more accurately by increasing the scatter in their forecasts for later series.

People were overconfident in their forecasts. Without OFB in the first session, overconfidence became even higher in the second session. People expected experience at the task, even in the absence of feedback, to improve their performance (Harvey & Fischer, 2005). As it did not do so, experience at the task produced an increase in overconfidence. However, provision of OFB in the first session disabused them of this notion. As a result, their overconfidence did not increase across sessions. Again, OFB continued to have an effect even once it had been removed: its influence can therefore be characterised as one of learning rather than as a product of incentivisation. Consistently with H_4, OFB lowered confidence, reduced bias, and improved calibration.

Analyses of inflation showed effects of OFB on both inflation judgments and confidence judgments but there were no significant interactions showing that these effects were greater later in the first session. In other words, there was no dose-response effect: more feedback was no more effective than less feedback. It appears that all that mattered was that participants received some feedback. We examine this issue in the next experiment.

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

3 Experiment 2

Experiment 1 showed beneficial effects of OFB but failed to show that more feedback was more beneficial. There are two possibilities. First, though the experiment had sufficient statistical power to reveal the main effect of feedback condition, it may not have had the power needed to pick up the three-way interaction. Second, it may be that a lot of feedback is indeed no more beneficial than just a little feedback. This somewhat counterintuitive notion receives support from the literature on skilled behaviour (Harvey, 2011). A number of studies using positioning tasks employed the same type of experimental design that we have adopted here: in a first (learning) session, groups performed the task under different feedback conditions; in a second (retention) session, all groups performed without feedback. We will briefly consider these studies.

During learning, Schmidt et al., (1989) gave different groups either OFB on every trial or summary OFB after every five, 10, or 15 trials; the summaries described the errors that they had made since they last received feedback. During learning, error was greater when feedback was less frequent but, at retention, this effect reversed: error was *lower* with less frequent feedback. During learning, Wulf and Schmidt (1989) provided one group with simple OFB on every trial and another group with simple OFB on two-thirds of the trials. Performance in the two groups was not significantly different during learning but those who had received *less* feedback performed better at retention. Winstein and Schmidt (1990) replicated this pattern of results in experiments in which feedback on all trials during learning was compared with feedback on 50% and 33% of trials. Schmidt and Bjork (1992) argued that these counterintuitive effects occur because making the task more difficult by reducing feedback causes people to put more effort into their task. This additional effort pays off once feedback has been removed in the retention session.

In the experiment that we report next, participants were randomly allocated into two sets of four groups. The two sets differed according to the type of feedback they received in the first session. Those in the first set received either no feedback ('no-feedback') or else simple OFB after every trial (i.e., ten times - '10-feedback'), after every other trial (i.e., five times - '5-feedback'), or after every five trials (i.e., twice - '2-feedback'). Those in the second set received either no feedback or else summary OFB after every trial ('10-feedback'), after every other trial ('5-feedback'), or after every five trials ('2-feedback'). No group received feedback in the second session.

Following the findings reviewed above, we expected inflation judgments during session one to be better with feedback than without it but to be no worse when feedback is given less often (H_6). Also, though studies in the literature on skill learning have not, as far as we are aware, directly compared effects of summary OFB (Schmidt et al., 1989) with simple OFB (Wulf & Schmidt, 1989), we expected judgments to be better in the former case (H_7). This was because summary OFB averaged over a number of trials provides information about the quality of more past judgments and is less subject to random noise than simple OFB.

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

3.1 Method

3.1.1 Participants

Four hundred and fifty-one participants (277 males) were recruited with the mean age of 26 years (SD = 13 years) via the online platform, Prolific.com. Data from two participants were excluded because their responses were incomplete. Each person was paid £1.00 for their participation and an additional £0.10 for every inflation judgment in the second session that was within 20% of the correct value. Data were collected from 23 December 2020 to 8 March 2021.

3.1.2 Design

This was a mixed design experiment with Session (first versus second) and Trial (1–10) as within-participant variables and Feedback Type (simple OFB versus summarised OFB) and Feedback Amount (no-feedback, 2-feedback, 5-feedback, 10-feedback) as between-participant variables. Participants were randomly allocated to one of these eight conditions.

3.1.3 Stimulus materials

Inflation data for 20 countries were identical to those used in Experiment 1. Information on the feedback page varied according to feedback condition. The simple OFB page was identical to that used in the first experiment: it showed the actual inflation rate and the participant's estimated inflation rate for the immediately preceding trial. The summarized OFB page showed the actual inflation rate averaged over one, two, or five previous trials (depending on the condition), the *averaged* inflation judgment averaged over one, two, or five previous trials (depending on the condition), and the difference between these two values. An example of a summarized OFB page is "Over the last two countries, the average inflation rate was 2.32% and your average forecast was 2.84%, so you overestimated by 0.52% on average".

3.1.4 Procedure

The procedure was identical to Experiment 1. The only differences were in the amount of OFB given (none, after every five trials, after every other trial, after every trial) and its type (simple versus summary).

3.2 Results

Participants' data were excluded when their inflation judgment for any of the 20 countries was beyond three standard deviations of the group mean judgment of that country. Implementation of this exclusion criterion resulted in data from 388 participants (241 males) with a mean age of 26 years (SD = 9 years) being entered into the analyses. Trial-by-trial data

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

of mean inflation judgments and mean confidence judgments for each of the eight groups are shown in Tables A2 and A3 of the Appendix, respectively. Means of these inflation judgments and their associated accuracy levels across all 20 countries in the experiment, together with corresponding values produced by two simple models, are given in Table A4 of the Appendix.

3.2.1 Forecasting performance

Measures of forecasting performance (absolute error, constant error, variable error) were derived in the same way as in the first experiment.

Sixteen one-sample t-tests showed that the mean CE values in both sessions of all eight conditions were significantly different from zero ($p < 0.001$). This provides further evidence consistent with H_1.

To examine the effects of OFB on inflation judgments (H_6 and H_7), we carried out four-way mixed ANOVAs on the raw inflation judgments, AE, CE, and VE with Feedback Type (simple OFB versus summarised OFB) and Feedback Amount (no-feedback, 2-feedback, 5-feedback, 10-feedback) as between-participant factors and with Session (one versus two) and Trial (1–10) as within-participant factors.

The analysis of inflation judgments (Figure 4, upper panel) showed only a main effect of Feedback Amount ($F(3, 380) = 7.16$, $p < 0.001$, *ges* = 0.0014). One-tailed post-hoc paired comparison tests revealed significant differences only between the no-feedback and the 2-feedback condition ($p = 0.007$), the 5-feedback condition ($p = 0.001$), and the 10-feedback condition ($p = 0.02$). A Scheffé test comparing inflation judgments in the no-feedback condition with those in all feedback conditions combined confirmed the beneficial effect of feedback ($p = 0.024$). These findings are consistent with H_6: feedback decreased inflation judgments but the effect did not depend on how much feedback was provided.

Analysis of AE (Figure 4, second panel) showed only a main effect of Feedback Amount ($F(3, 380) = 3.49$, $p = 0.02$, *ges* = 0.0008). However, no post-hoc comparisons reached significance.

Analysis of CE (Figure 4, third panel) revealed a main effect of Feedback Amount ($F(3, 380) = 7.16$, $p < 0.001$, *ges* = 0.0019). One-tailed paired comparison tests showed significant differences only between the no-feedback condition and the 2-feedback condition ($p = 0.01$), the 5-feedback condition ($p = 0.001$), and the 10-feedback condition ($p = 0.04$). A Scheffé test comparing the no-feedback condition with the three feedback conditions combined confirmed the beneficial effect of feedback ($p = 0.005$). As Figure 4 shows, CE was, on average, 33% higher in the no-feedback group. These findings are consistent with H_6. However, there was no evidence that summarized OFB was more beneficial than simple OFB (H_7). Finally, there a marginally significant interaction between Feedback Amount, Session, and Trial ($F(26.46, 3351.60) = 1.51$, $p = 0.05$, *ges* = 0.0054): while there was a small but constant difference between the no-feedback condition and the feedback conditions in Session 1, that difference became larger over Session 2.

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

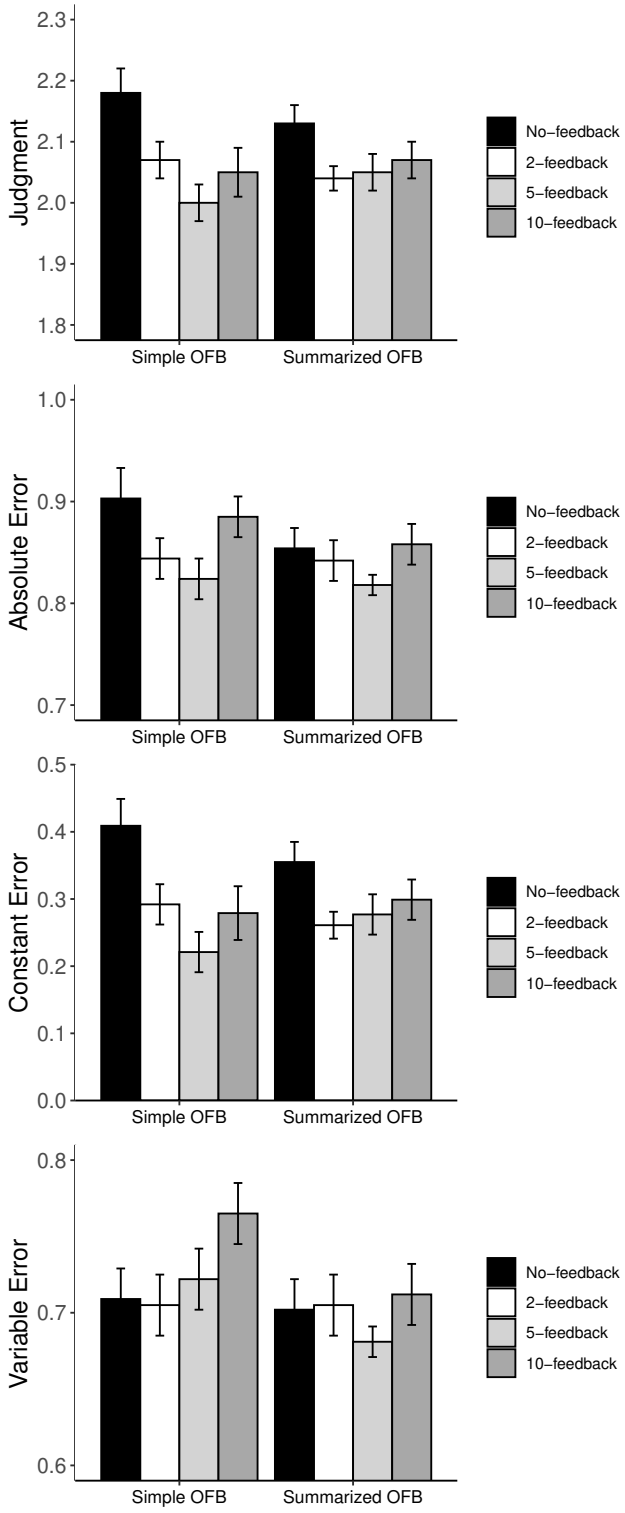

FIGURE 4: Experiment 2: Mean Judgment (upper panel), Absolute Error (second panel), Constant Error (third panel), and Variable Error (lower panel) in each group, all shown with standard error bars.

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

Analysis of VE (Figure 4, lower panel) revealed only a main effect of Trial (F = (7.78, 2954.88) = 2.06, p =0.04, *ges* = 0.0023). This arose because VE showed an inverted U-shaped trend across the 10 trials (Table A2). Although the overall effect of feedback amount was not significant, we expected that our Experiment 1 finding that simple feedback on every trial produces a higher level of VE than no feedback would be replicated. A Bonferroni post-hoc test showed that this comparison was indeed significant (p = 0.03), with VE in the 10-feedback simple OFB condition (0.76) higher than VE in the no-feedback groups (0.71).

3.2.2 Confidence in forecasts

Measures of confidence in forecasts (bias, calibration score) were derived in the same way as in the first experiment.

Sixteen one-sample t-tests showed that the mean Bias scores in both sessions of all eight conditions were significantly different from zero (p < 0.001). This provides further evidence consistent with H_3.

Analysis of confidence levels (Figure 5, upper panel) failed to show a significant main effect of Feedback Amount (p = 0.15). However, as we had an *a priori* expectation that feedback would reduce levels of confidence (H_4), we used a Scheffé test to compare confidence in the no-feedback condition with that in all feedback conditions combined. This showed that feedback did indeed reduce people's confidence in their inflation judgments (p < 0.001). There was also a main effect of Trial (F (8.26, 3139.56) = 3.11, p = 0.001, *ges* = 0.0011). This arose because high initial confidence dropped over the early part of each session (Table A3).

Analysis of bias (Figure 5, middle panel) showed a marginally significant main effect of Feedback Amount (F (3, 380) = 2.21, p = 0.087, *ges* = 0.0028). However, as we had an *a priori* expectation that feedback would reduce Bias (H_4), we used a Scheffé test to compare confidence in the no-feedback condition with that in all feedback conditions combined. This showed that feedback did indeed reduce Bias in judgments of inflation (p = 0.007): on average, bias scores in the no-feedback groups were 4% greater than those in the feedback groups. There was also a main effect of Feedback Type (F (1, 380) = 5.72, p = 0.02, *ges* = 0.0024). This arose because bias scores were 0.05 greater when participants received summarised OFB than when they received simple OFB.

Analysis of calibration scores (Figure 5, lower panel) showed a marginal main effect of Feedback Amount (F (3, 380) = 2.21, p = 0.086), *ges* = 0.0043). Given that we expected feedback to improve calibration (H_4), we used a Scheffé test to compare calibration score in the no-feedback groups with that in the feedback groups. This confirmed that calibration was better with feedback (p < 0.001): on average, calibration scores in the no-feedback groups were 3% lower than those in the feedback groups.

There was a main effect of Trial (F (8.52, 3238.74) = 2.92, p = 0.002, *ges* = 0.0027) and interactions between Feedback Type, Feedback Amount, and Session (F (3, 380) = 3.68, p = 0.01, *ges* = 0.0014) and between Trial, Feedback Type, Feedback Amount, and Session (F

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

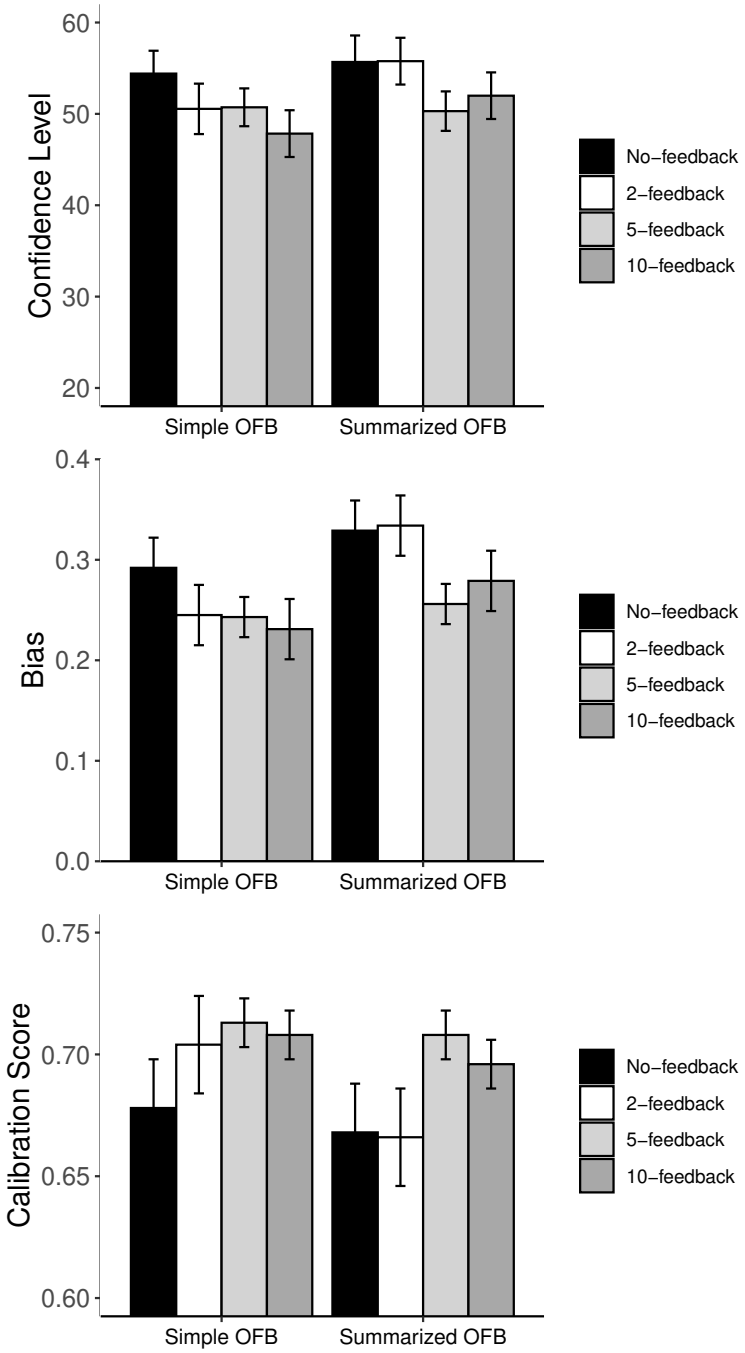

FIGURE 5: Experiment 2: Mean Confidence level (upper panel), Bias (middle panel), and Calibration score (lower panel) in each group, all shown with standard error bars.

(25.46, 3225.06) = 1.72, p = 0.01, *ges* = 0.0047). With simple OFB, calibration remained fairly constant over sessions with a mean of 0.71; with summarised OFB, it again remained fairly constant but with a slightly lower mean value (0.69); with no OFB, calibration was lower still with a mean value of 0.67 but showing some improvement across trials that itself varied over the two sessions.

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

We again examined the evidence for a hard-easy effect (H_5). Figure 6 shows separate graphs for the hard-easy effect in the no-feedback group and in the three feedback groups combined. For the no-feedback group, the regression revealed that Bias = 0.55 – 0.01 Difficulty (Adj R^2 = 0.99; F (1, 18) =1274.18, p < 0.001). And for the feedback groups, the regression showed that Bias = 0.51 – 0.01 Difficulty (Adj R^2 = 0.99; F (1, 18) = 1807.85, p < 0.001). Thus, as in the first experiment, it is clear that the effect of feedback was to lower the intercept while leaving the slope unchanged.

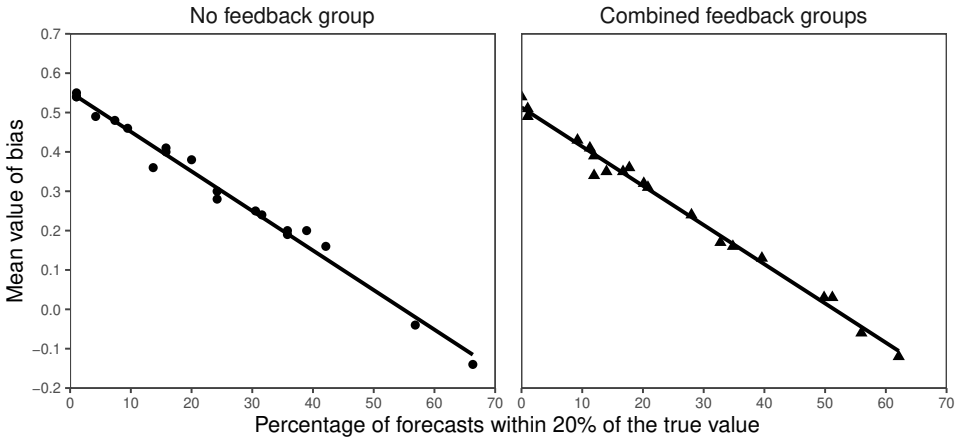

FIGURE 6: Experiment 2: Bias plotted against forecast difficulty in the no-feedback group (left panel) and the feedback groups (right panel).

3.3 Discussion

The results confirmed that inflation judgments are better with feedback provision but they were no worse when feedback was given less often (H_6). Specifically, these judgments were lower and, hence, less biased when feedback was given in the first session than when it was not given. However, giving feedback more frequently did not produce any greater reduction in constant error. Furthermore, against our expectations (H_7), simple OFB was just as beneficial as summarized OFB. There was no sign that this beneficial effect of feedback was reduced after feedback had been removed in the second session: this implies that its effect was again due to learning rather than incentivisation.

In contrast to the first experiment, provision of OFB had no *overall* effect on variable error. However, the comparison of the no-feedback group with the 10-feedback group revealed a significant effect in the simple OFB conditions but not in the summarized OFB conditions. Thus we replicated our finding from the first experiment but the effect that we found did not generalise to the summarized OFB conditions. Why was this? In the first

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

experiment, feedback merely specified the actual inflation rate along with the participant's judged inflation rate. In the current experiment, this was true only in the simple OFB conditions. In the summarized OFB conditions, participants were also explicitly given the difference between the actual and judged inflation rate (e.g., "you overestimated by 0.52% on average"). If this made feedback easier to process and thereby reduced judgment noise for participants in the summarized feedback conditions, VE would have been lowered such that it became close to that observed in the no-feedback group.

Absolute error reflects both constant and variable error. In our first experiment, feedback had beneficial effects on the former but detrimental effects on the latter; the effects of OFB on these two types of error cancelled each other out and, as a result, feedback had no overall effect on AE. In the current experiment, there was a main effect of the amount of feedback but post-hoc comparisons failed to reach significance. However, the opposite effects of providing feedback on every trial on CE and VE in the simple OFB conditions were still statistically significant and their contributions to AE cancelled each other out just as they did in the first experiment (Figure 4). The overall effect of feedback amount may reflect lower AE values in the 2-feedback and 5-feedback conditions than in the no-feedback and 10-feedback conditions. However, we have no statistical evidence to support this interpretation.

We expected feedback to lower confidence, reduce bias, and improve calibration (H_4) and we found that it did have these effects. However, as with inflation judgments, providing more feedback did not have a more beneficial effect. In addition, overconfidence was greater when participants received summarized OFB than when they received simple OFB. We assume that this occurred because people given summarised OFB erroneously believed that such feedback would benefit their performance more than people given simple OFB expected that that type of feedback would benefit their performance. This implies that people have an imperfect internal model of the factors that affect their performance (c.f., Harvey and Fischer, 2005).

There were some main and interactive effects of Trial. These were generally small and, though not specifically anticipated, most of them make sense in the context of the other results. The difference in CE between feedback and no-feedback groups increased over the experiment: provision of feedback reduced CE and kept it relatively low whereas, in the absence of feedback, CE drifted higher over time. With experience at the task, confidence dropped after the first trial. Analysis of calibration scores showed high level interactions which, while not easy to interpret, appear to reflect the changeable quality of both inflation judgments and confidence judgments in the absence of feedback but their relative stability in its presence. Finally, in all groups, VE scores were relatively low at the beginning and end of each session but somewhat higher middle. People may start sessions by making relatively cautious judgments close to the mean of the presented series, become more adventurous as they suspect that some series must show trends, and then revert to caution when experience of more stimulus series convinces them that none do.

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

4 General discussion

We first consider inflation forecasts and the effect that OFB had on them. We then discuss confidence in those forecasts and the effect that OFB had on it. Next, we assess implications of our findings for consumers and for the way in which surveys of inflation expectations are used by central banks and other agencies. We then outline some limitations of our work and, finally, provide a brief concluding section.

4.1 Inflation forecasts

In both experiments, inflation forecasts were too high. This is consistent with previous research on lay people's expectations of inflation and may arise because they find it easier to bring to mind past examples of high price rises and use these as a basis for their forecasts (Bruine de Bruin et al., 2011). However, in our experiments, forecasts were not memory-based; there was no need to recall past price rises. Instead, people were given past values of inflation from previous years to base their forecasts on. (These are typically available for respondents in experts' surveys, such as the SPF, but not for respondents in lay surveys, such as the MSC.) Furthermore, in our experiments, people made forecasts for countries other than their own: their recall of particular price rises in their own countries should not have been relevant (though it may still have had an influence). Another explanation for over-forecasting of inflation, at least in experts, has been that forecasters may be penalised less for over-forecasting than for under-forecasting (e.g., Capistrán and Timmermann, 2009). However, in our experiments, performance-related incentives were not asymmetrical. Thus it is more likely that, as Niu and Harvey (2021) suggest, people expect inflation to be more likely to rise than fall. This may be a risk-amplification effect of using the availability heuristic: media coverage of potential and actual rises in inflation is more extensive than corresponding coverage of falls in inflation.

Outcome feedback reduced over-forecasting. As far as we are aware, this is the first time that such an effect has been demonstrated for time series forecasting. In past studies, successive trials have required people to make forecasts from the same time series and so provision of outcome feedback has been unavoidable (in order to provide the latest point for the next forecast). As a result, previous studies were only able to use an OFB condition as a baseline against which to assess the effectiveness of more sophisticated types of feedback (e.g., Goodwin & Fildes, 1999; Remus et al., 1996). In contrast, by presenting forecasters with a different time series on each trial, we were able to compare the OFB condition with a no-feedback condition. Furthermore, the lack of any significant reduction in the effect of feedback after it had been removed in the second session is not consistent with the feedback having its effect by providing an incentive for better performance; instead, it is best explained by a learning effect (Annett, 1969).

However, this learning does not appear to have been a slow incremental process of the sort that is characteristic of much skill acquisition. In the first experiment, there was no

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

evidence that the advantage of the feedback group over the no-feedback group increased over trials. In the second experiment, groups receiving feedback produced less biased inflation judgments than those not receiving feedback but there was no evidence that groups receiving more feedback were less biased than those receiving less feedback. This suggests that provision of just some information was sufficient for bias to be reduced. It appears that OFB transmitted knowledge to forecasters and that this knowledge needed to be transmitted just once or twice to be effective.

Recently, there has been greater appreciation that the overall error in judgments depends not just on biases but also on the scatter or noise that they contain (Kahneman et al., 2021). Previous studies of judgmental forecasting have shown that different factors influence these two error components and that the same factor can affect them in different ways (e.g., Harvey & Bolger, 1996). The present study provides further evidence of this. In the first experiment, OFB reduced bias (CE) but increased scatter (VE); as a result, it had no effect on overall error (AE). This effect was replicated in the simple OFB conditions of the second experiment: VE was again higher and CE was lower when feedback was provided on every trial than when no feedback was provided and, consequently, there was no resultant effect of feedback on AE. In the summarized OFB conditions, feedback did not increase VE: we attributed this to the fact feedback given in these conditions included an explicit comparison of the participant's judgment and the outcome, thereby reducing one contributor to judgment noise.

4.2 Confidence in forecasts

On average, people were 28% overconfident in their inflation forecasts in the first experiment and 28% overconfident in them in the second experiment. This degree of overconfidence in forecasts is higher than that obtained by Wright (1982) and Ronis and Yates (1987). It is unlikely that this difference arose because our participants were asked to assess the likelihood that an outcome would appear within an interval whereas theirs were asked to estimate the likelihood that one of two possible answers was correct. This is because Hansson et al. (2008) also found low overconfidence, comparable to that reported by Ronis and Yates (1987), when they asked people to estimate the probability that an outcome would fall within a specified interval. It is more likely that the difference relates to the fact that Wright (1982), Ronis and Yates (1987) and Hansson et al. (2008) all studied memory-based judgments whereas we examined time-series forecasting.

Outcome feedback reduced overconfidence by 9% in the first experiment and by 5% in the second one. As our task involved a sequence of judgments to related items, these findings are consistent with the conclusions of our earlier review: OFB reduces biases in confidence judgments when those judgments are made to related items (Benson & Önkal, 1992; Subbotin, 1996; Winman & Juslin, 1993) but not when they are made to unrelated items (Keren, 1988; Subbotin, 1996). We could attribute the only exception to this generalisation, as indeed did its author (Zakay, 1992), to an incentive effect of OFB that

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

increased people's attention to critical aspects of the task. However, the beneficial influence of OFB on confidence judgments in our task is not attributable to such an effect: this is because there was no evidence of a significant reduction in that influence after feedback was withdrawn in the second session. Instead the benefits of OFB must have arisen from some type of learning effect.

What was the nature of this learning? Figures 3 and 6 indicate that feedback had no effect whatsoever on the hard-easy effect: the intercept of the regression of bias on to task difficulty was reduced by feedback but its slope remained the same. A finding such as this can be interpreted within Ferrell's (1994) model of calibration. According to his approach, the skill of good calibration depends on two separate abilities that combine to produce accurate probability judgments: base-rate identification and discriminability. The intercepts of the regressions shown in Figures 3 and 6 depend on an ability to identify the base-rate of the occurrence of an event: in our task, the ability to determine the average likelihood of a forecast being within 20% of the correct value. The slopes of the regressions depend on an ability to discriminate between more likely and less likely events: in our task, the ability to discriminate between series for which producing a forecast is so difficult that it is unlikely that it will be within 20% of the outcome and series for which producing a forecast is much easier so that it is much more likely that it will be within 20% of the outcome. Without this second ability, people will, on average, judge all series close to the identified base rate and, as a result, they will be more overconfident when judging more difficult series. Our findings show that OFB improved base-rate identification but had no effect on discriminability.

Provision of OFB significantly improved calibration but, even with this feedback, the mean calibration score was only 0.70. This is not especially impressive: as we have seen, a uniform judge who always considers that the forecast is as likely as not to be within 20% of the outcome would obtain a calibration score of 0.75. This poor calibration after provision of OFB is not especially surprising given that, as we have just seen, feedback improved people's ability to judge the mean difficulty of forecasting but did not improve their ability to discriminate between series that were harder to forecast and those that were easier to forecast.

4.3 Implications

Lay people tend to expect inflation to be higher than it turns out to be. This is assumed to have a number of effects on their economic behaviour. For example, they are likely to bring forward their purchasing of durable goods; this will, in turn, produce higher demand for those goods and increase their prices. In other words, expectations of higher inflation increase inflation. This is why central banks use surveys of lay expectations of inflation to help them forecast inflation. It follows that more realistic inflation expectations should decrease inflation. We have shown that training with OFB can produce more realistic expectations. However, one would not want to provide such training only to survey respondents as their

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

more realistic expectations would not then be representative of the population from whom they were drawn.

Central banks carry out surveys of lay and professional forecasters for different reasons. Whereas lay expectations *feed into* the forecasting process because those expectations are assumed to influence economic behaviour and hence inflation, professional expectations represent one way of *providing* forecasts. Thus, there are no downsides to improving inflation forecasts made by those responding to professional surveys; there are no concerns about altering the degree to which professional forecasters responding to surveys are representative of the population of professional forecasters as a whole. Training professional survey respondents to produce better forecasts could reduce their tendency to over-forecast inflation (Ang et al., 2007). It would be relatively easy to implement as professional surveys (unlike lay surveys) already use the approach that we adopted here and provide respondents with past values of inflation when asking them to forecast future ones.

Although survey respondents have been asked to assess aleatory uncertainty associated with future inflation (e.g., assess the likelihood that inflation will exceed 4% next year), they have not been asked to assess epistemic uncertainty (e.g., assess the likelihood that your estimate of next year's inflation is within 20% of what it turns out to be). However, there is an argument that central banks should put more weight on survey expectations that respondents produce with greater certainty. If those who expect inflation to be high are not as certain of their expectations as those who expect it to be low, more emphasis should be given to the latter view when data are aggregated. This approach would work better when respondents are better calibrated. Training with OFB could be helpful here.

4.4 Limitations and suggestions for future work

Our experiments demonstrated the effectiveness of training with OFB. We showed that such training remains effective even after feedback has been withdrawn. However, a question remains about how long it would remain effective. If the interval between the first and second sessions of our experiments had been extended to hours or days, would the advantage of training still persist?

Training using guidance has been found to be more effective than training with OFB when people have little experience in performing a task and when tasks are complex (Holding & Macrae, 1964; Macrae & Holding, 1965). Also, within the MCPL tradition, guidance (task information) has been found to be more effective than feedback (Balzer et al., 1989). Guidance provides advance information about a task that can be incorporated into the knowledge used to perform it. Our finding that people benefit from provision of OFB but are insensitive to the amount provided suggests that this feedback provides them with useful knowledge about their task; once they have obtained this knowledge, they receive no additional advantage from being given it again. In other words, just like guidance, OFB provides task information but, unlike guidance, it is provided after rather than before performance. This line of thinking suggests that it would be worth investigating

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

whether guidance is as effective for improving inflation forecasting as training with OFB. It would be relatively simple to implement: even providing a warning that the average survey respondent overestimates inflation by some percentage could be effective.

We presented our participants with inflation rates for the years 2009–2018 and required them to forecast the inflation rate for 2019. In fact, 14 out of the 20 countries showed a fall in CPI inflation from 2018 to 2019; inflation rates in one country are not independent of those in other countries. Although participants were not informed of the base rates of different types of inflation change, those receiving OFB may, over the first experimental session, have slowly come to realize (correctly) that inflation was more likely to fall than rise. However, this would imply that inflation judgments would be lower at the end of the first session in the feedback groups but not in the no-feedback groups. In fact, as we have seen, inflation judgments showed a U-shaped function over the 10 trials in each session in both feedback and no-feedback groups in Experiment 1 (Table A1) and no trend over the 10 trials in any group in Experiment 2 (Table A2). These data are not consistent with cross-series dependencies informing participants' judgments. (To confirm this, simulated series could be used to manipulate the base rate systematically to determine whether it has any effect on participants' judgments.)

We have shown that OFB decreases over-forecasting of inflation. There may be other manipulations that have a similar effect. Inflation rates are typically small numbers ranging between -1.00 and 2.00. Laypeople, especially those with low numeracy, may not find it easy to process such numbers. Indeed, less educated people have been found to have higher expectations of future inflation rates, possibly because they round up their expectations in order to express them as whole numbers (Bryan and Venkatu, 2001a). This suggests that recoding inflation by, say, multiplying inflation rates by 100, might reduce over-forecasting by such people (c.f., Sedlmeier, 2000). However, such recoding would result in values that people would find difficult to relate to their everyday experience of price changes. An increase of one percent is easily interpretable as an extra cent in every dollar whereas an increase of 100 units in every 10,000 units is not so easy to process.

5 Concluding comments

In the past, many approaches to improving decision making have been examined in a variety of domains that have been found to facilitate performance without bringing it very close to optimal. Here we have shown for the first time that one of the simplest of these approaches, provision of OFB, is effective in reducing biases in time series forecasting and in the confidence that people have in their forecasts. However, there was little effect on overall error in forecasts because OFB increased the noise or scatter in forecasts. In practice, this should not detract from the advantages of using OFB because, unlike bias, noise or scatter can be reduced by averaging over a number of different judges (Surowiecki, 2004) or over a number of judgments from the same individual (Herzog & Hertwig, 2014).

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

References

Adams, P. A., & Adams, J. K. (1958). Training in confidence-judgments. *The American Journal of Psychology*, *71*(4), 747–751. https://doi.org/10.2307/1420334.

Andersson, P., & Rakow, T. (2007). Now you see it now you don't: The effectiveness of the recognition heuristic for selecting stock. *Judgment and Decision Making, 2*(1), 29–39. http://journal.sjdm.org/jdm06162.pdf.

Ang, A., Bekaert, G. & Wei, M. (2007). Do macro variables, asset markets, or surveys forecast inflation best? *Journal of Monetary Economics, 54*(4)*,* 1163–1212. https://doi.org/10.1016/j.jmoneco.2006.04.006

Annett, J. (1969). *Feedback and human behaviour: The effects of knowledge of results, incentives and reinforcement on learning and performance.* London: Penguin Books.

Ayton, P., Önkal, D., & McReynolds, L. (2011). Effects of ignorance and information on judgments and decisions. *Judgment and Decision Making*, *6*(5), 381–391. http://journal.sjdm.org/11/rh7/rh7.html.

Balzer, W. K., Doherty, M. E. and O'Connor, R. Jr. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin, 106*(3), 410–33. https://psycnet.apa.org/doi/10.1037/0033-2909.106.3.410.

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics*, *55*(4), 412–428. https://doi.org/10.3758/BF03205299.

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. https://doi.org/10.1016/0169-2070(92)90066-I.

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*(2*),* 127–151. https://doi.org/10.1016/j.obhdp.2006.07.001.

Borges, B., Goldstein, D. G., Ortmann, A., & Gigerenzer, G. (1999). Can ignorance beat the stock market? In G. Gigerenzer, P. M. Todd and the ABC Research Group (Eds.) *Simple heuristics that make us smart,* Oxford: Oxford University Press. (pp 59-72).

Brehmer, B. (1978). Response consistency in probabilistic inference tasks. *Organizational Behavior and Human Performance, 22*(1), 103-115. https://doi.org/10.1016/0030-5073(78)90008-9.

Brehmer, B. (1980). In one word: Not from experience. *Acta Psychologica*, *45*(1–3), 223–241. https://doi.org/10.1016/0001-6918(80)90034-7.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Brown, S. D., & Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology, 58*(1), 49 – 67. https://psycnet.apa.org/doi/10.1016/j.cogpsych.2008.09.

002.

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. https://doi.org/10.1016/j.joep.2011.07.002.

Bryan, M. F., & Venkatu, G. (2001a). The demographics of inflation opinion surveys. *Federal Reserve Bank of Cleveland, Economic Commentary*, 10.15.2001. https://core.ac.uk/download/pdf/6888515.pdf.

Bryan, M. F., & Venkatu, G. (2001b). The curiously different inflation perspectives of men and women. *Federal Reserve Bank of Cleveland, Economic Commentary*. 11.01.2001. https://www.clevelandfed.org/en/newsroom-and-events/publications/economic-commentary/economic-commentary-archives/2001-economic-commentaries/ec-20011101-the-curiously-different-inflation-perspectives-of-men-and-women.aspx.

Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, *19*(1), 7–42. https://doi.org/10.1023/A:1007850605129.

Capistrán, C. & Timmermann, A. (2009). Disagreement and biases in inflation expectations. *Journal of Money, Credit and Banking, 41*(2–3), 365–396. https://doi.org/10.1111/j.1538-4616.2009.00209.x.

Chang, W., Chen, E., Mellers, B. & Tetlock, P. (2016). Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments, *Judgment and Decision Making, 11*(5), 509–526. https://www.sas.upenn.edu/baron/journal/16/16511/jdm16511.pdf.

Cohen, J. & Dearnley, E. J. (1962). Skill and judgment of footballers in attempting to score goals: A study of psychological probability. *British Journal of Psychology, 53*(1), 71–88. https://doi.org/10.1111/j.2044-8295.1962.tb00815.x.

Cohen, J., Dearnaley, E. J., & Hansel, C. E. M. (1956). Risk and hazard: Influence of training on the performance of bus drivers. *Journal of the Operational Research Society*, *7*(3), 67–82. https://doi.org/10.1057/jors.1956.13.

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications*. Academic press.

Deane, D. H., Hammond, K. R., & Summers, D. A. (1972). Acquisition and application of knowledge in complex inference tasks. *Journal of Experimental Psychology*, *92*(1), 20–26. https://psycnet.apa.org/doi/10.1037/h0032162

Edwards, W & Fasolo, B. (2001). Decision Technology. *Annual Review of Psychology, 52,* 581–606. https://doi.org/10.1146/annurev.psych.52.1.581.

Ferrell, W.R. (1994). Discrete subjective probabilities and decision analysis: Elicitation, calibration and combination. In G. Wright and P. Ayton (Eds.) *Subjective probability*. New York: Wiley (pp 411–451).

Fiedler, K., Kareev, Y., Avrahami, J., Beier, S., Kutzner, F., & Hütter, M. (2016). Anomalies in the detection of change: When changes in sample size are mistaken for changes in

proportions. *Memory & Cognition, 44*(1), 143–161. https://doi.org/10.3758/s13421-015-0537-z.

Fildes, R., & Petropoulos, F. (2015). Improving forecast quality in practice. *Foresight: The International Journal of Applied Forecasting*, *36* (Winter), 5–12. https://orca.cardiff.ac.uk/84332.

Fischer, G. W. (1982). Scoring-rule feedback and the overconfidence syndrome in subjective probability forecasting. *Organizational Behavior and Human Performance*, *29*(3), 352–369. https://doi.org/10.1016/0030-5073(82)90250-1.

Fischer, I., & Harvey, N. (1999). Combining forecasts: What information do judges need to outperform the simple average? *International Journal of Forecasting*, *15*(3), 227–246. https://doi.org/10.1016/S0169-2070(98)00073-9.

Fischhoff, B., & MacGregor, D. (1982). Subjective confidence in forecasts. *Journal of Forecasting*, *1*(2), 155–172. https://doi.org/10.1002/for.3980010203.

Frensch, P. A., & Funke, J. (eds.). (1995). *Complex Problem Solving: The European Perspective.* Hillsdale, NJ: Erlbaum.

Georganas, S., Healy, P. J., & Li, N. (2014). Frequency bias in consumers' perceptions of inflation: An experimental study. *European Economic Review*, *67*, 144–158. https://doi.org/10.1016/j.euroecorev.2014.01.014.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence. *Psychological Review*, *98*(4), 506–528. https://psycnet.apa.org/doi/10.1037/0033-295X.98.4.506.

Goldberg, L. R. (1968). Simple models or simple processes? Some research on clinical judgments. *American Psychologist*, *23*(7), 483–496. https://psycnet.apa.org/doi/10.1037/h0026206.

Goodwin, P., & Fildes, R. (1999). Judgmental forecasts of time series affected by special events: Does providing a statistical forecast improve accuracy? *Journal of Behavioral Decision Making*, *12*(1), 37–53. https://doi.org/10.1002/(SICI)1099-0771(199903)12:1<37::AID-BDM319>3.0.CO;2-8.

Hammond, K. R. (1971). Computer graphics as an aid to learning. *Science*, *172*(3986), 903–908. https://doi.org/10.1126/science.172.3986.903.

Hammond, K. R., & Boyle, P. J. (1971). Quasi-rationality, quarrels and new conceptions of feedback. *Bulletin of the British Psychological Society, 24,* 103–113. https://psycnet.apa.org/record/1972-03933-001.

Hammond, K. R., & Summers, D. A. (1972). Cognitive control. *Psychological Review*, *79*(1), 58–67. https://psycnet.apa.org/doi/10.1037/h0031851.

Hammond, K. R., Summers, D. A., & Deane, D. H. (1973). Negative effects of outcome-feedback in multiple-cue probability learning. *Organizational Behavior and Human Performance*, *9*(1), 30–34. https://doi.org/10.1016/0030-5073(73)90034-2.

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 Ex-*

perimental Psychology: Learning, Memory, and Cognition, 34(5), 1027–1042. https://psycnet.apa.org/doi/10.1037/a0012638.

Harries, C., & Harvey, N. (2000). Taking advice, using information and knowing what you are doing. *Acta Psychologica*, *104*(3), 399–416. https://doi.org/10.1016/S0001-6918(00)00038-X.

Harvey, N. (1994). Relations between confidence and skilled performance. In G. Wright & P. Ayton (Eds.), *Subjective probability,* New York: John Wiley & Sons. (pp. 321–352).

Harvey, N. (1995). Why are judgments less consistent in less predictable task situations? *Organizational Behavior and Human Decision Processes*, *63*(3), 247–263. https://doi.org/10.1006/obhd.1995.1077.

Harvey, N. (2007). Use of heuristics: Insights from forecasting research. *Thinking & Reasoning*, *13*(1), 5–24. https://doi.org/10.1080/13546780600872502.

Harvey, N. (2011). Learning judgment and decision making from feedback. In M. K. Dhami, A. Schlottmann, and M. R. Waldmann (Eds.), *Judgment and decision making as a skill: Learning, development, and evolution*, Cambridge: Cambridge University Press. (pp. 406–464).

Harvey, N., & Bolger, F. (1996). Graphs versus tables: Effects of data presentation format on judgemental forecasting. *International Journal of Forecasting*, *12*(1), 119–137. https://doi.org/10.1016/0169-2070(95)00634-6.

Harvey, N., Ewart, T., & West, R. (1997). Effects of data noise on statistical judgement. *Thinking & Reasoning*, *3*(2), 111–132. https://doi.org/10.1080/135467897394383.

Harvey, N., & Fischer, I. (2005). Development of experience-based judgment and decision making: The role of outcome feedback. In T. Betsch & S. Haberstroh (Eds.) *The routines of decision making.* Mahwah, NJ: Lawrence Erlbaum Associates. (pp. 119–137).

Herzog, S. M. & Hertwig, R. (2014). Harnessing the wisdom of the inner crowd. *Trends in Cognitive Sciences, 18*(10), 504–506. https://doi.org/10.1016/j.tics.2014.06.009.

Holding, D. H., & Macrae, A. W. (1964). Guidance, restriction and knowledge of results. *Ergonomics*, *7*(3), 289–295. https://doi.org/10.1080/00140136408930748.

Holzworth, R. J., & Doherty, M. E. (1976). Feedback effects in a metric multiple-cue probability learning task. *Bulletin of the Psychonomic Society*, *8*(1), 1–3. https://doi.org/10.3758/BF03337054.

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes*, *57*(2), 226–246. https://doi.org/10.1006/obhd.1994.1013.

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

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin*, *134*(3), 404–426. https://psycnet.apa.org/doi/10.1037/0033-2909.134.3.404.

Keren, G. (1988). On the ability of monitoring non-veridical perceptions and uncertain knowledge: Some calibration studies. *Acta Psychologica, 67*(2), 95–119. https://doi.org/10.1016/0001-6918(88)90007-8.

Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica*, *77*(3), 217–273. https://doi.org/10.1016/0001-6918(91)90036-Y.

Kerstholt, J. H. (1996). *Dynamic decision making.* Wageningen: Ponsen & Looijen B.V.

Kim, I. & Kim, M. (2009). Irrational bias in inflation forecasts. *Munich Personal RePEc Archive (MPRA) paper No 16447*. https://mpra.ub.uni-muenchen.de/16447.

Klayman, J. (1988). On the how and why (not) of learning from outcomes. In B. Brehmer, & C. J. B. Joyce (Eds.), Human judgement: The SJT view (*Advances in psychology*, Vol. 54). Amsterdam: Elsevier North-Holland. (pp. 115–162).

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. https://doi.org/10.1016/j.ijforecast.2006.03.007.

Lei, C., Lu, Z. & Zhang, C. (2015). News on inflation and the epidemiology of inflation expectations in China. *Economic Systems, 39*(4), 644–653. https://doi.org/10.1016/j.ecosys.2015.04.006.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*(2), 255–275. https://psycnet.apa.org/doi/10.1037/0033-2909.125.2.255.

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance*, *26*(2), 149–171. https://doi.org/10.1016/0030-5073(80)90052-5.

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

Macrae, A. W. & Holding, D. H. (1965). Guided practice in direct and reversed serial tracking. *Ergonomics, 8*(4), 487–92. https://doi.org/10.1080/00140136508930830.

McClelland, A. G., & Bolger, F. (1994). The calibration of subjective probability: Theories and models 1980–94. In: Wright G and Ayton P (eds). *Subjective probability.* Wiley: Chichester. (pp. 453–482).

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S.E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E. & Tetlock, P. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science, 25*(5), 1106–1115. https://doi.org/10.1177/0956797614524255.

Nisbett, R. E. (Ed.) (1993). *Rules for reasoning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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

3 September 2021: https://doi.org/10.1016/j.ijforecast.2021.07.009.

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. https://psycnet.apa.org/doi/10.1037/1082-989X.8.4.434.

Önkal, D., & Muradoğlu, G. (1995). Effects of feedback on probabilistic forecasts of stock prices. *International Journal of Forecasting*, *11*(2), 307–319. https://doi.org/10.1016/0169-2070(94)00572-T.

Osman, M. (2010). Controlling uncertainty: A review of human behaviour in complex dynamic environments. *Psychological Bulletin, 136*(1), 65–86. https://psycnet.apa.org/doi/10.1037/a0017815.

Pachur, T., & Biele, G. (2007). Forecasting from ignorance: The use and usefulness of recognition in lay predictions of sports events. *Acta Psychologica*, *125*(1), 99–116. https://doi.org/10.1016/j.actpsy.2006.07.002.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Cyrino Oliveira, F. L., De Baets, S., Dokumentov, A. . . . Ziel, F. (2022). Forecasting: Theory and practice. *International Journal of Forecasting*, Published online 20 January.

Remus, W., O'Conner, M., & Griggs, K. (1996). Does feedback improve the accuracy of recurrent judgmental forecasts? *Organizational Behavior and Human Decision Processes*, *66*(1), 22–30. https://doi.org/10.1006/obhd.1996.0035.

Ronis, D. L., & Yates, J. F. (1987). Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior and Human Decision Processes*, *40*(2), 193–218. https://doi.org/10.1016/0749-5978(87)90012-4.

Russo, J. E., & Schoemaker, P. J. (1992). Managing overconfidence. *Sloan Management Review*, *33*(2), 7–17. https://sloanreview.mit.edu/article/managing-overconfidence.

Schmitt, N., Coyle, B. W., & King, L. (1976). Feedback and task predictability as determinants of performance in multiple cue probability learning tasks. *Organizational Behavior and Human Performance*, *16*(2), 388–402. https://doi.org/10.1016/0030-5073(76)90023-4.

Schmitt, N., Coyle, B. W., & Saari, B. B. (1977). Types of task information feedback in multiple-cue probability learning. *Organizational Behavior and Human Performance*, *18*(2), 316–328. https://doi.org/10.1016/0030-5073(77)90033-2.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*(4), 207–218. https://doi.org/10.1111/j.1467-9280.1992.tb00029.x.

Schmidt, R. A., Young, D. E., Swinnen, S., & Shapiro, D. C. (1989). Summary knowledge of results for skill acquisition: Support for the guidance hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(2), 352. https://psycnet.apa.org/doi/

10.1037/0278-7393.15.2.352.

Schuh, S. (2001). An evaluation of recent macroeconomic forecast errors. *New England Economic Review*, *January/February*, 35–36. https://www.bostonfed.org/publications/new-england-economic-review/2001-issues/issue-number-1-januaryfebruary-2001/an-evaluation-of-recent-macroeconomic-forecast-errors.aspx.

Sedlmeier, P. (2000). How to improve statistical thinking: Choose the task representation wisely and learn by doing. *Instructional Science, 28*(3), 227–262. https://doi.org/10.1023/A:1003802232617.

Serwe, S., & Frings, C. (2006). Who will win Wimbledon? The recognition heuristic in predicting sports events. *Journal of Behavioral Decision Making*, *19*(4), 321–332. https://doi.org/10.1002/bdm.530.

Sharp, G. L., Cutler, B. L., & Penrod, S. D. (1988). Performance feedback improves the resolution of confidence judgments. *Organizational Behavior and Human Decision Processes*, *42*(3), 271–283. https://doi.org/10.1016/0749-5978(88)90001-5.

Sniezek, J. A. (1990). A comparison of techniques for judgmental forecasting by groups with common information. *Group & Organization Studies*, *15*(1), 5–19. https://doi.org/10.1177/105960119001500102.

Sterman, J. D. (1989). Misperceptions of feedback in dynamic decision making. *Organizational Behavior and Human Decision Processes*, *43*(3), 301–335. https://doi.org/10.1016/0749-5978(89)90041-1.

Stone, E. R., & Opel, R. B. (2000). Training to improve calibration and discrimination: The effects of performance and environmental feedback. *Organizational Behavior and Human Decision Processes*, *83*(2), 282–309. https://doi.org/10.1006/obhd.2000.2910.

Subbotin, V. (1996). Outcome feedback effects on under-and overconfident judgments (general knowledge tasks). *Organizational Behavior and Human Decision Processes, 66*(3), 268–276. https://doi.org/10.1006/obhd.1996.0055.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations.* New York: Random House.

Thurstone, L. L. (1926). The scoring of individual performance. *Journal of Educational Psychology*, *17*(7), 446–457. https://psycnet.apa.org/doi/10.1037/h0075125.

Todd, F. J., & Hammond, K. R. (1965). Differential feedback in two multiple-cue probability learning tasks. *Behavioral Science*, *10*(4), 429–435. https://doi.org/10.1002/bs.3830100406.

Winman, A., & Juslin, P. (1993). Calibration of sensory and cognitive judgments: Two different accounts. *Scandinavian Journal of Psychology*, *34*(2), 135–148. https://doi.org/10.1111/j.1467-9450.1993.tb01109.x.

Winstein, C. J., & Schmidt, R. A. (1990). Reduced frequency of knowledge of results enhances motor skill learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(4), 677–691. https://psycnet.apa.org/doi/10.1037/0278-7393.16.4.

677.

Wright, G. (1982). Changes in the realism and distribution of probability assessments as a function of question type. *Acta Psychologica*, *52*(1–2), 165–174. https://doi.org/10.1016/0001-6918(82)90033-6.

Wulf, G., & Schmidt, R. A. (1989). The learning of generalized motor programs: Reducing the relative frequency of knowledge of results enhances memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 748–757. https://psycnet.apa.org/doi/10.1037/0278-7393.15.4.748.

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance*, *30*(1), 132–156. https://doi.org/10.1016/0030-5073(82)90237-9.

Yates, J. F., & Curley, S. P. (1985). Conditional distribution analyses of probabilistic forecasts. *Journal of Forecasting*, *4*(1), 61–73. https://doi.org/10.1002/for.3980040110.

Zakay, D. A. N. (1992). The influence of computerized feedback on overconfidence in knowledge. *Behaviour & Information Technology*, *11*(6), 329–333. https://doi.org/10.1080/01449299208924354.

Appendix

This appendix includes tables of summarized data showing inflation judgments and confidence judgments on each trial in each session in Experiment 1 (Table A1) and Experiment 2 (Tables A2 and A3), together with a comparison of error scores in each experiment with two simple algorithmic models (Table A4). Full data for each participant in both experiments are available at https://osf.io/k8vx9.

TABLE A1: Experiment1: Means and standard deviations (in parentheses) of inflation judgments and confidence judgments.

		Trial										
Inflation judgments		1	2	3	4	5	6	7	8	9	10	mean
Feedback	Session 1	2.34(1.52)	1.96(1.11)	1.95(1.48)	1.67(1.29)	2.15(1.45)	1.86(1.05)	1.50(1.04)	1.94(1.47)	2.36(1.73)	2.29(1.71)	2.00(0.42)
	Session 2	2.44(1.50)	1.77(1.19)	2.02(1.24)	1.77(1.52)	2.13(1.47)	1.65(1.15)	1.82(1.21)	1.75(1.16)	2.22(1.32)	2.28(1.43)	1.98(0.34)
	mean	2.39(0.98)	1.87(0.79)	1.99(0.99)	1.72(0.92)	2.14(0.92)	1.76(0.76)	1.66(0.80)	1.85(0.88)	2.29(1.01)	2.28(1.03)	1.99(0.22)
No-feedback	Session 1	2.64(1.71)	2.11(1.27)	2.14(1.51)	2.37(1.51)	1.81(1.04)	1.84(1.14)	2.13(1.21)	2.09(1.40)	2.26(1.51)	2.29(1.20)	2.17(0.40)
	Session 2	2.34(1.46)	1.96(1.48)	1.98(1.25)	2.14(1.22)	2.15(1.52)	2.32(1.47)	2.28(1.53)	2.08(1.23)	2.04(1.52)	2.10(1.06)	2.14(0.42)
	mean	2.49(1.15)	2.04(0.97)	2.06(1.04)	2.25(0.96)	1.98(0.94)	2.08(0.95)	2.20(1.10)	2.09(0.88)	2.15(0.97)	2.20(0.62)	2.15(0.28)
Confidence judgments												
Feedback	Session 1	53.04(20.31)	51.72(19.51)	49.39(21.28)	47.09(22.04)	46.37(20.30)	49.76(23.17)	49.76(24.86)	45.09(24.87)	44.50(24.73)	49.89(24.40)	48.66(18.91)
	Session 2	47.78(22.37)	46.74(23.01)	47.33(22.65)	47.35(21.40)	46.20(23.53)	48.43(21.73)	46.85(24.46)	46.43(22.70)	48.65(23.33)	49.17(23.98)	47.49(20.20)
	mean	50.41(17.05)	49.23(19.64)	48.36(20.61)	47.22(19.88)	46.28(20.44)	49.10(20.75)	48.30(23.43)	45.76(22.59)	46.58(22.45)	49.53(22.52)	48.08(19.19)
No-feedback	Session 1	55.80(22.67)	58.76(22.24)	55.15(25.51)	56.12(21.98)	54.71(19.52)	59.39(24.95)	59.17(25.34)	52.71(23.32)	57.27(21.64)	52.88(24.90)	56.20(18.91)
	Session 2	60.17(21.39)	58.61(21.83)	59.41(23.12)	62.37(22.46)	58.27(23.35)	61.41(21.79)	56.05(24.98)	63.66(21.25)	60.49(22.94)	59.41(24.88)	59.99(19.15)
	mean	57.99(20.58)	58.68(20.43)	57.28(21.80)	59.24(19.55)	56.49(18.49)	60.40(20.34)	57.61(22.14)	58.18(19.88)	58.88(20.29)	56.15(22.54)	58.09(18.50)

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

TABLE A2: Experiment 2: Means and standard deviations (in parentheses) of inflation judgments.

		Trial										
Simple OFB		1	2	3	4	5	6	7	8	9	10	mean
No-feedback	Session 1	2.19(1.43)	2.07(1.40)	1.80(0.99)	1.95(1.15)	2.48(1.59)	1.94(1.07)	1.86(1.14)	2.23(1.70)	2.26(1.23)	2.97(1.89)	2.17(0.38)
	Session 2	2.38(1.72)	1.99(1.36)	2.08(1.21)	2.19(1.45)	2.27(1.60)	2.16(1.43)	2.02(1.36)	2.26(1.40)	2.44(1.75)	2.14(1.40)	2.19(0.39)
	mean	2.28(1.25)	2.03(1.02)	1.94(0.85)	2.07(0.97)	2.38(1.26)	2.05(0.84)	1.94(0.94)	2.24(1.12)	2.35(1.05)	2.55(1.19)	2.18(0.26)
2-feedback	Session 1	2.33(1.61)	1.91(1.04)	2.32(1.47)	1.94(1.26)	2.12(1.45)	1.94(1.16)	1.93(1.13)	2.20(1.47)	2.16(1.04)	2.02(1.36)	2.09(0.36)
	Session 2	1.93(1.26)	2.12(1.42)	2.16(1.33)	2.02(1.18)	2.02(1.26)	2.16(1.62)	1.87(1.08)	2.31(1.54)	1.86(1.08)	2.02(1.11)	2.05(0.41)
	mean	2.13(0.87)	2.01(0.88)	2.24(0.83)	1.98(0.86)	2.07(0.94)	2.05(0.99)	1.90(0.72)	2.25(0.99)	2.01(0.74)	2.02(0.87)	2.07(0.20)
5-feedback	Session 1	2.13(1.31)	1.99(1.07)	2.06(1.25)	1.94(1.34)	1.98(1.32)	1.92(1.25)	1.85(1.30)	2.09(1.38)	2.08(1.29)	1.84(1.36)	1.99(0.39)
	Session 2	1.83(1.23)	1.89(1.74)	1.82(1.23)	1.99(1.20)	2.04(1.37)	2.53(1.35)	1.91(1.39)	2.12(1.58)	1.68(1.17)	2.21(1.49)	2.00(0.41)
	mean	1.98(0.93)	1.94(0.88)	1.94(0.89)	1.97(0.80)	2.01(0.95)	2.22(0.77)	1.88(0.95)	2.11(1.07)	1.88(0.72)	2.02(1.08)	2.00(0.20)
10-feedback	Session 1	2.13(1.48)	1.96(1.40)	1.86(1.41)	2.09(1.44)	2.38(1.64)	2.09(1.39)	1.89(1.61)	2.30(1.69)	1.90(1.19)	2.09(1.38)	2.07(0.44)
	Session 2	2.12(1.62)	2.02(1.31)	1.85(1.06)	1.98(1.39)	1.64(0.76)	2.02(1.33)	2.58(1.50)	2.07(1.30)	2.03(1.28)	2.06(1.47)	2.04(0.40)
	mean	2.13(0.89)	1.99(0.85)	1.86(0.87)	2.04(1.02)	2.01(0.94)	2.06(0.84)	2.23(1.13)	2.19(1.00)	1.97(0.79)	2.07(1.02)	2.05(0.25)
Summarized OFB												
No-feedback	Session 1	2.09(1.34)	2.33(1.51)	2.58(1.48)	1.96(1.28)	1.97(1.41)	2.16(1.11)	1.99(1.15)	2.11(1.16)	1.66(1.11)	2.11(1.60)	2.10(0.35)
	Session 2	2.40(1.54)	2.19(1.40)	2.15(1.43)	1.76(1.37)	2.27(1.42)	2.02(1.40)	2.34(1.38)	2.21(1.53)	2.14(1.44)	2.13(1.33)	2.16(0.40)
	mean	2.25(0.92)	2.26(0.97)	2.36(0.80)	1.86(0.87)	2.12(0.96)	2.09(0.79)	2.17(0.92)	2.16(0.92)	1.90(1.00)	2.12(1.03)	2.13(0.19)
2-feedback	Session 1	1.92(1.27)	1.93(1.07)	2.11(1.46)	1.98(1.29)	2.31(1.49)	1.88(1.12)	1.94(1.09)	1.95(1.27)	2.22(1.44)	2.23(1.49)	2.05(0.35)
	Session 2	1.87(1.02)	1.97(1.20)	2.05(1.25)	1.97(1.11)	1.97(1.20)	2.02(1.30)	2.07(1.38)	1.93(1.37)	2.14(1.35)	2.26(1.55)	2.03(0.35)
	mean	1.89(0.87)	1.95(0.69)	2.08(0.95)	1.98(0.81)	2.14(0.95)	1.95(0.85)	2.00(0.80)	1.94(1.06)	2.18(0.91)	2.25(1.15)	2.04(0.18)
5-feedback	Session 1	2.18(1.31)	2.28(1.40)	2.10(1.30)	1.88(1.38)	2.01(1.52)	1.92(1.33)	2.03(1.11)	2.11(1.41)	2.06(1.34)	1.80(1.03)	2.04(0.35)
	Session 2	1.91(1.19)	1.97(1.34)	2.18(1.47)	2.17(1.40)	1.87(1.23)	2.03(1.39)	2.13(1.44)	2.02(1.50)	2.28(1.53)	2.09(1.32)	2.07(0.36)
	mean	2.05(0.77)	2.12(0.95)	2.14(0.88)	2.02(0.96)	1.94(0.90)	1.98(1.02)	2.08(0.87)	2.07(1.06)	2.17(1.03)	1.95(0.84)	2.05(0.21)
10-feedback	Session 1	2.25(1.46)	2.41(1.54)	2.00(1.22)	2.14(1.62)	1.98(1.09)	1.80(1.31)	2.00(1.22)	2.15(1.45)	1.94(1.42)	2.19(1.25)	2.09(0.34)
	Session 2	2.06(1.10)	2.00(1.38)	1.86(1.31)	2.12(1.22)	2.35(1.51)	2.15(1.62)	1.97(1.31)	2.14(1.55)	2.10(1.43)	1.88(1.14)	2.06(0.33)
	mean	2.16(1.01)	2.20(1.04)	1.93(0.91)	2.13(1.10)	2.16(0.93)	1.98(0.94)	1.98(0.88)	2.15(1.15)	2.02(0.98)	2.03(0.94)	2.07(0.20)

Note. Table A4 summarizes the accuracy levels of judgmental forecasts in Experiment 1 and Experiment 2 and of forecasts from two simple algorithmic models a) Linear regression model: each forecast was produced by regressing the 10 historical data points for a country and using the model to predict the inflation rate for 2019, b) Mean model: each forecast was produced by extracting the mean of the 10 years' historical inflation data points for a country and using it as the forecast for 2019. (The mean of actual inflation rates in 2019 for the 20 countries was 1.77% with the SD of 1.72%.)

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press

TABLE A3. Experiment2: Means and standard deviations (in parentheses) of confidence judgments.

		Trial										
Simple OFB		1	2	3	4	5	6	7	8	9	10	mean
No-feedback	Session 1	51.54(21.52)	55.04(19.17)	51.85(21.23)	53.60(17.76)	53.83(20.71)	52.21(20.86)	57.17(21.85)	56.94(20.94)	54.92(23.28)	54.38(21.81)	54.15(17.20)
	Session 2	54.38(22.41)	56.88(21.00)	56.81(19.65)	54.17(21.12)	54.81(20.32)	53.17(20.52)	53.54(21.81)	55.81(20.67)	50.96(22.57)	56.25(21.30)	54.68(18.13)
	mean	52.96(19.51)	55.96(17.66)	54.33(18.62)	53.89(16.90)	54.32(19.08)	52.69(18.41)	55.35(20.15)	56.38(19.12)	52.94(20.57)	55.31(20.37)	54.41(17.29)
2-feedback	Session 1	50.00(21.11)	50.91(21.70)	50.33(21.44)	51.20(20.33)	50.70(20.61)	49.63(21.57)	49.04(21.94)	48.74(22.27)	52.26(22.00)	50.70(22.99)	50.35(19.08)
	Session 2	52.17(20.26)	50.78(22.31)	51.89(21.45)	47.65(20.38)	51.80(19.76)	48.24(23.68)	50.30(20.65)	49.52(22.15)	53.37(21.70)	51.83(20.18)	50.76(18.99)
	mean	51.09(19.03)	50.85(20.21)	51.11(19.38)	49.42(18.58)	51.25(19.24)	48.93(21.25)	49.67(20.19)	49.13(21.12)	52.82(20.98)	51.26(20.88)	50.55(18.69)
5-feedback	Session 1	53.60(17.52)	50.55(16.70)	51.00(17.84)	52.74(15.39)	47.83(18.08)	50.45(17.98)	50.96(16.82)	51.96(20.20)	51.79(20.02)	53.57(19.74)	51.44(14.01)
	Session 2	52.09(17.82)	50.15(17.31)	48.81(16.32)	49.64(19.49)	47.43(17.61)	50.17(18.76)	50.43(19.52)	50.81(19.11)	49.00(19.85)	51.51(19.07)	50.00(15.39)
	mean	52.84(15.65)	50.35(14.10)	49.90(14.74)	51.19(15.51)	47.63(16.04)	50.31(16.73)	50.69(15.70)	51.38(17.02)	50.39(18.56)	52.54(18.22)	50.72(14.20)
10-feedback	Session 1	52.21(20.42)	47.70(20.51)	46.81(18.96)	47.98(21.11)	47.47(20.51)	46.74(20.50)	49.49(21.02)	46.72(22.80)	45.47(22.07)	46.58(22.39)	47.72(16.35)
	Session 2	49.47(21.68)	50.14(20.87)	46.72(23.42)	45.21(20.33)	49.14(22.03)	48.98(22.99)	48.53(20.39)	47.98(21.73)	47.72(21.02)	45.84(22.40)	47.97(18.71)
	mean	50.84(18.20)	48.92(17.97)	46.77(18.76)	46.59(16.85)	48.30(17.74)	47.86(18.82)	49.01(19.40)	47.35(20.06)	46.59(19.78)	46.21(20.53)	47.84(16.79)
Summarized OFB												
No-feedback	Session 1	60.43(17.47)	55.43(22.20)	55.55(21.31)	52.28(21.34)	52.60(22.75)	54.64(18.34)	56.38(20.89)	52.06(20.27)	54.51(20.91)	52.96(22.07)	54.68(17.57)
	Session 2	58.11(23.40)	57.55(25.00)	58.55(25.46)	55.91(23.87)	56.85(24.35)	56.96(24.09)	55.47(24.99)	57.70(26.19)	57.11(27.34)	52.47(27.19)	56.67(23.61)
	mean	59.27(18.94)	56.49(22.08)	57.05(21.07)	54.10(19.96)	54.72(22.09)	55.80(18.69)	55.93(21.39)	54.88(21.26)	55.81(22.66)	52.71(22.51)	55.68(19.85)
2-feedback	Session 1	57.56(21.51)	55.15(21.44)	53.23(24.01)	56.65(22.08)	54.37(21.84)	54.63(21.85)	54.37(21.41)	55.33(22.75)	55.12(22.41)	54.12(22.26)	55.05(18.53)
	Session 2	58.40(22.15)	55.92(21.28)	56.35(20.53)	57.06(21.66)	57.31(21.25)	55.25(21.54)	54.60(22.15)	57.46(20.86)	57.65(22.17)	54.81(23.24)	56.48(18.95)
	mean	57.98(20.20)	55.54(19.45)	54.79(19.74)	56.86(20.73)	55.84(19.60)	54.94(18.71)	54.48(19.66)	56.39(20.97)	56.38(21.28)	54.46(20.73)	55.77(18.44)
5-feedback	Session 1	53.70(17.82)	50.44(17.11)	51.11(18.02)	47.26(17.38)	50.52(19.99)	51.07(18.03)	51.28(18.97)	51.11(19.66)	53.19(19.04)	47.57(18.24)	50.73(15.07)
	Session 2	52.09(17.55)	50.11(19.48)	49.91(20.35)	52.06(22.07)	48.98(20.85)	51.87(21.38)	47.04(20.09)	48.56(21.80)	49.56(19.94)	48.50(21.62)	49.87(17.58)
	mean	52.90(15.67)	50.28(16.46)	50.51(17.20)	49.66(17.92)	49.75(18.11)	51.47(17.99)	49.16(16.98)	49.83(17.88)	51.37(17.67)	48.04(18.37)	50.30(15.90)
10-feedback	Session 1	55.94(19.10)	52.00(17.06)	50.04(18.43)	47.12(21.01)	52.61(21.08)	51.06(21.94)	49.94(23.54)	53.78(24.74)	50.71(21.53)	50.14(21.82)	51.33(16.51)
	Session 2	51.96(21.47)	51.78(22.18)	50.49(22.67)	51.29(22.32)	53.02(22.92)	55.41(23.78)	54.37(23.71)	52.37(24.22)	54.16(22.10)	51.65(23.82)	52.65(20.76)
	mean	53.95(16.96)	51.89(17.02)	50.26(18.33)	49.21(19.95)	52.81(20.56)	53.24(21.40)	52.16(21.38)	53.08(21.89)	52.43(20.62)	50.89(21.34)	51.99(18.22)

TABLE A4: Means and standard deviations (in parentheses) of different types of inflation forecast over 20 countries.

Inflation forecasts	Judgment level	Absolute error	Constant error	Variable error
Linear regression model	1.61(1.64)	0.90(1.15)	-0.17(1.47)	0.91(1.13)
Mean model	2.31(1.86)	0.75(0.67)	0.53(0.86)	0.65(0.54)
Judgments in Experiment 1	2.07(1.38)	0.88(0.86)	0.29(1.19)	0.73(0.68)
Judgments in Experiment 2	2.07(1.36)	0.85(0.83)	0.30(1.15)	0.71(0.65)

https://doi.org/10.1017/S1930297500009050 Published online by Cambridge University Press