ELSEVIER

International Journal of Forecasting 20 (2004) 29–39

international journal of forecasting

www.elsevier.com/locate/ijforecast

The effects of feedback on judgmental interval predictions

Fergus Bolger*, Dilek Önkal-Atay

Faculty of Business Administration, Bilkent University, Ankara, Turkey

Abstract

The majority of studies of probability judgment have found that judgments tend to be overconfident and that the degree of overconfidence is greater the more difficult the task. Further, these effects have been resistant to attempts to 'debias' via feedback. We propose that under favourable conditions, provision of appropriate feedback should lead to significant improvements in calibration, and the current study aims to demonstrate this effect. To this end, participants first specified ranges within which the true values of time series would fall with a given probability. After receiving feedback, forecasters constructed intervals for new series, changing their probability values if desired. The series varied systematically in terms of their characteristics including amount of noise, presentation scale, and existence of trend. Results show that forecasts were initially overconfident but improved significantly after feedback. Further, this improvement was not simply due to 'hedging', i.e. shifting to very high probability estimates and extremely wide intervals; rather, it seems that calibration improvement was chiefly obtained by forecasters learning to evaluate the extent of the noise in the series.

Keywords: Judgmental forecasting; Calibration; Feedback; Overconfidence; Confidence intervals

1. Introduction

Research into the accuracy (or 'calibration') of judgmental forecasts has produced consistent findings of overconfidence (see Lichtenstein, Fischhoff, & Phillips, 1982; Arkes, 2001, for reviews). To give a specific example, if a weather forecaster predicts a 70% chance of rain on 100 different days, then rain should be observed on 70 of those days for the forecaster to be perfectly calibrated—typically what is found is that rain will occur on fewer than 70 days. A related finding is that the extent of observed overconfidence depends on task difficulty such that the more difficult the task, the greater the degree of overconfidence—this is known as the 'hard–easy' effect (Lichtenstein et al., 1982). Good calibration may be observed for tasks with an average proportion correct around 75%, while *underconfidence* may be observed for tasks with proportions correct greater than 75%. However, most tasks have a lower proportion correct than 75%, thus contributing to the overall finding of *overconfidence* (see, e.g. Suantak, Bolger, & Ferrell, 1996, for a review).

The currently most favoured explanation for overconfidence is that it occurs due to some kind of error in mapping true feelings of confidence on to the required response scale (see Ayton & McClelland, 1997, for a review). It should therefore follow that response error and overconfidence could be reduced

* Corresponding author. School of Psychology, University of Leicester, University Road, Leicester LE1 7RH, UK. Tel.: +44-116-252-2155; fax: +44-116-252-2067.

E-mail address: fmib1@le.ac.uk (F. Bolger).

0169-2070/$ - see front matter © 2003 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/S0169-2070(03)00009-8

by practice with *feedback*[1], in other words, people should be able to learn the appropriate mappings. However, studies of probability judgment which make use of general-knowledge questions have typically failed to significantly reduce overconfidence with feedback. These tests normally require judgments for *unrelated* events (see, e.g. Keren, 1991) and answers which are deliberately selected to be counter-intuitive (see, for example, Gigerenzer, Hoffrage, & Kleinbolting, 1991)—neither of these conditions are ideal for learning. In contrast, a forecasting task offers better opportunities for learning from feedback, as probability judgments have to be made repeatedly for related events which have not been selected to be misleading. In fact, there is some evidence to suggest that overconfidence may not be as high in forecasting tasks (Wright & Ayton, 1986). Also, excellent calibration has been reported with experts in weather forecasting and in the game of bridge; their performance attributed in part to the availability of consistent and timely outcome feedback (see Keren, 1987; Murphy & Winkler, 1984, respectively).

However, this past work on forecasting has mainly involved discrete judgmental probability forecasts (i.e. the probability that the predicted event will occur). A potentially simpler structure for eliciting and communicating the uncertainties inherent in the forecasting process is provided by interval predictions (i.e. forecasts offering a minimum and a maximum bound within which a future value is expected to lie with a specified probability). Such forecasts depict an intuitively meaningful and uncomplicated format for both the providers and users of predictions, whilst conveying detailed information about future expectations (Önkal-Atay, Thomson, & Pollock, 2002).

For interval estimation tasks, a judge would be considered as perfectly calibrated if intervals given a 95% confidence coefficient actually contain the true event on 95% of occasions. However, as for discrete judgments, for non-forecasting tasks, overconfidence is typically found with intervals. Thus, for instance, 95% confidence interval assessments contain the true event only 60–70% of the time. In other words, confidence is generally too high relative to interval width. Lichtenstein and Fischhoff (1980) found that 40% of realized values actually fall outside the 98% intervals specified by participants in their study. Striking examples in business settings are provided by Russo and Schoemaker (1992), who asked managers to estimate confidence intervals for uncertain quantities in their own areas of expertise (e.g. petroleum, banking, advertising, data processing, etc.). The resulting hit rates were 21–22% for 50% confidence intervals, 36–58% for 90% intervals, and 20–42% for 95% confidence intervals.

Again, as for discrete judgments, there is some evidence that overconfidence in interval judgments with general knowledge items appear to resist improvement through practice with feedback. For instance, Alpert and Raiffa (1982) used general knowledge questions and after ten answers, urged subjects to ‘Spread Those Extreme Fractiles! Be honest with yourselves! Admit what you don’t know!’ (p. 301). In the follow-up set of ten questions, subjects showed only a modest improvement. Further, Plous (1995) studied intervals given by groups (for general knowledge items) and found that ‘overconfidence persisted in the face of explicit warnings, instructions to expand interval widths, and extended group discussion’ (p. 451). However, there also exists some evidence that calibration of interval judgments may indeed improve with outcome feedback (O’Connor & Lawrence, 1989; Roth, 1993). The O’Connor and Lawrence (1989) study is of particular interest because it investigates *interval* judgment in a *forecasting* task, thus containing two features that we anticipate should permit well-calibrated forecasts to be learned. The study to be reported here—where we aim to demonstrate that, under favourable conditions, overconfidence can be eliminated by the provision of appropriate feedback—consequently bears certain similarities to O’Connor and Lawrence’s (1989) study.

O’Connor and Lawrence (1989) presented 33 students, who had no prior experience of time-series forecasting or confidence intervals, with just one time series. Each student received a *different* time series and made seven forecasts, receiving *outcome* feedback immediately after each forecast. The time

[1] This assumes that the error is correctable rather than, for instance, ‘cognitive noise’, which is not correctable. The response-error models are not clear on this point but, given the observance of good calibration in a number of studies, it seems there must be a sizeable proportion of correctable error otherwise the response-error models cannot be a full explanation for miscalibration.

series used were *real* series selected from the M-competition (Makridakis et al., 1982). In our study we wished to improve on this design in the following ways:

1. increase the number of participants so as to improve statistical power (we used 139 participants);
2. use participants with at least some basic knowledge of time-series forecasting and confidence intervals (our participants were management students completing a one-semester forecasting course and had recently attended classes on these topics);
3. use constructed rather than real series in order to control the series characteristics;
4. require each forecaster to make forecasts for a range of different time series with varying characteristics in order to simulate more realistic conditions for both forecasting and learning (in our study the participants made forecasts for 96 different series in all); and
5. investigate the role of *calibration* feedback rather than outcome feedback (i.e. summarize the percentage of hits over a number of trials at a particular level of confidence).

The first four items are self-explanatory. The last item needs to be expanded upon and it is to this issue we turn to next.

A number of theorists have identified several different types of feedback. For example, a basic distinction has been made between *outcome* and *cognitive* feedback (Todd & Hammond, 1965). The former refers to the so-called 'knowledge of results' such as the true value of some variable that was forecast; whereas the latter refers to information about relations such as that between the outcome and one's prediction (i.e. forecast error) or between the outcome and features of the series (such as its trend or variability). Balzer, Doherty, and O'Connor (1989) go on to distinguish three components of cognitive feedback: *task information* (TI), *cognitive information* (CI) and *functional validity information* (FVI). With reference to multiple cue probability learning (MCPL) tasks, TI reflects relations between cues and criterion, for example, the extent to which it is possible to estimate the criterion value from cues and the intercorrelations between cues. CI refers to relations between a person's judgment and the cues such as the level and variability of judgment across cues in a MCPL task. Finally, FVI details relations between a judge's cognitive strategy and the task, for instance, the correlation between judged and actual values of the criterion in MCPL tasks. Balzer, Sulsky, Hammer, and Sumnar (1992) found that only feedback about TI produced significant improvements in performance on an MCPL task—similar results have also been found by Balzer, Hammer, Sumner, Birchenough, Martens, and Raymark (1994).

Benson and Önkal (1992) made a somewhat similar distinction to Balzer et al. (1989)—but in relation to judgmental forecasting rather than MCPL tasks—between *performance* and *environmental* feedback. In particular, performance feedback involves providing information about the accuracy of judgment in general, such as information on the *calibration* of judgmental forecasts, and can be regarded as feedback of a form of FVI. Environmental feedback, on the other hand, involves providing information about the event to be predicted, such as its predictability from series characteristics, a form of TI. Stone and Opel (2000) argue that in order to produce improvements in *calibration* of probability judgments it is necessary to provide performance feedback/FVI, and *not* environmental feedback/TI, although the latter is required to produce improvements in *discrimination* (i.e. separating instances when a target event will or will not occur). In a probability judgment task, Stone and Opel (2000) found support for their view and further found that overconfidence was *greater* after provision of environmental feedback/TI. Other researchers have also found calibration to improve after provision of performance feedback/FVI (e.g. Adams & Adams, 1958; Benson & Önkal, 1992; Bornstein & Zickafoose, 1999; Lichtenstein & Fischhoff, 1980; Oskamp, 1962; Önkal & Muradoğlu, 1995; Sharp, Cutler, & Penrod, 1988). For this reason, the primary form of feedback in our probabilistic forecasting task is performance feedback/FVI in the form of information about forecasters' calibration. In addition, forecasters also receive some environmental feedback/TI *incidentally*, as is described below. It is worth noting that, analyzing point forecasts only, Remus, O'Connor, and Griggs (1996) have found performance improvements via TI feedback relative to performance feedback given in the form of MAPE scores.

Cybernaticists (e.g. Weiner, 1948) were probably the first to ascribe to feedback a central role in behaviour. They proposed that feedback constitutes information about behaviour which is then used to *control* future behaviour. This idea has subsequently been developed and elaborated. For example, Powers (1973) argues that feedback is information about behaviour which can be compared against a cognitive standard or 'reference condition'. Differences between current behaviour and the reference condition result in changes in behaviour so as to reduce the mismatch. This 'error correction' process is used to achieve a system's behavioral *goals* in the context of small variations in the environment (which Powers refers to as 'disturbances'). Note that the idea of a reference condition implies a mental representation of the ideal desired goal state against which behavior can be compared. Powers (1973) also proposes that there is a *hierarchy of goals* such that at a low level we may try to reduce, say, error in prediction but at a higher level try to fulfil our goals of mastery over the environment or self-actualization.

Let us now try to apply these notions to the current study. The reference condition, at a primary goal level, we suggest is a representation of the amount of variability in the series. The feedback indicates how well the behavioral output—which is a statement regarding this variability in terms of an interval of a particular width for a particular level of confidence—actually corresponds to this representation across a range of environmental disturbances, which are the different characteristics of the time series presented. In our task, as already noted, there are two kinds of feedback: (1) environmental feedback/TI focusing on the way in which properties of the stimulus series varies (for example, that there are no long-term negative or damped trends in the stimulus sets); and (2) calibration feedback (the 'hit rate' or percent of true values of the series falling within a set of intervals of a given confidence level). The first of these two types of feedback is implicit and the second explicit, but both allow, at minimum, modification (or control) of the behavioural output (the interval judgment). However, as already reviewed, the literature suggests that the latter form of feedback is the most effective for improving calibration.

Now with respect to the proposed hierarchy of goals, we suggest that a higher level goal than accurately capturing the variance of series by means of one's confidence intervals is a *communicative* one: to inform others (in our case primarily the experimenter) of the uncertainty in the series, and thus in the forecast itself. This is in line with Yaniv and Foster's (1995, 1997) idea that interval forecasts are a trade-off between accuracy and informativeness. They suggest that people may sacrifice accuracy (i.e. insuring that the true value falls within the interval, which can be attained by giving very wide intervals) for the sake of informativeness (i.e. giving more precise estimates, best attained by assessing rather narrow intervals). For example, stating that the first trans-Atlantic flight occurred some time between the years 1800 and 2000 is highly likely to contain the true answer, but is unlikely to be considered very useful by a potential user of this estimate. In contrast, giving the interval 1920–1930 is likely to be considered a much more useful estimate even if it does not actually contain the true value. Communication is thus seen as an essential part of the forecasting process.

Yaniv and Foster (1995, 1997) go further and suggest that the accuracy-informativeness trade-off is a reason for observed overconfidence in interval judgments as the tendency to want to make intervals more informative produces pressure to make them narrower. Studies of interval judgments often insist on 90, 95 or even 99% confidence intervals which generally require very wide, and correspondingly uninformative, intervals thus the push towards narrower intervals creates overconfidence. In our study, we permit the forecasters to select their own confidence level, thus both accuracy and informativeness can be attained by reducing stated confidence levels whilst also narrowing the intervals.

2. Research hypotheses

2.1. Calibration and overconfidence

Following the results of the existing literature outlined above, our first hypothesis involves the pervasive overconfidence revealed in judgmental predictions:

Hypothesis 1: Interval forecasts will initially manifest overconfidence but this will be signifi-

> cantly reduced as forecasters receive feedback about their performance.

In the spirit of Powers' cybernetic model we propose that our forecasters will have a hierarchy of goals which they will try to satisfy using the available feedback. Their primary goal will be to match the width of intervals to their representation of the variability of the series in order to achieve good calibration:

> *Hypothesis 2*: Improvement in calibration due to feedback will be the result of forecasters being better able to reflect in their confidence intervals the variability in the time series.

2.2. Accuracy vs. informativeness

As we have seen, Yaniv and Foster (1995, 1997) found that people have a tendency to prefer informativeness to accuracy when providing confidence intervals. That is, narrow intervals may be considered more useful (i.e. more informative) for decision making than wide ones, but they are less likely to contain the true value (i.e. less accurate), all else being equal. Yaniv and Foster argue that this is one reason why people tend not to increase their ranges sufficiently with practice. In accordance with these arguments, we anticipate that a secondary goal of our forecasters will be to reduce the widths of their confidence intervals in order for them to be more informative. In order for them to do this and satisfy their primary goal of good calibration they must therefore simultaneously reduce the confidence levels they choose for their intervals:

> *Hypothesis 3*: Over sessions there will be a decrease in the width of the judgment intervals provided and an accompanying decrease in the confidence levels selected.

3. Method

3.1. Participants

A total of 139 third-year business students at Bilkent University, Turkey completed the experiment towards extra credit in a forecasting course (157 started but 18 did not complete all four sessions). The gender ratio was approximately equal.

3.2. Materials

A total of 96 52-week time-series graphs were used to elicit one-period-ahead forecasts. The last four values of the displayed series were also presented in tabular form next to the graph. Constructed series were used, and the participants were told that they showed the values of real Turkish stocks with undisclosed stock names and time periods.

The series varied in terms of the mean value of the stocks (three stock price levels: low, medium, high), degree of first-order autocorrelation (four levels: approximately 0.6, 0.3, 0 or -0.3), amount of noise (two levels: low and high), trend (two levels: positive linear trend and no trend), and scale (two levels: scaled-up 50%, not scaled-up) (see Fig. 1 for three example graphs). There were also three different levels of mean stock price which, when combined factorially with the other features of the series described above, results in the 96 different series. These 96 series were semi-randomly allocated to the three experimental sessions—that is to say that a complete factorial arrangement was maintained within each session for all attributes apart from mean level, with approximately equal numbers of the three mean levels in each session.

The parameters were all selected to reflect the behavior of actual Turkish stock price series at the time the experiment was conducted. For example, given the high inflation rate, Turkish stocks tended not to display any long-term negative trends, and were more likely to show positive than negative autocorrelation.

3.3. Procedure

The participants were tested in four groups of approximately equal size (four different sections of the third-year course in Forecasting). The experiment lasted four sessions, as described below.

***Session 1*:** Participants were asked to make one-step-ahead probabilistic interval forecasts for each of 32 graphically presented time-series. At the outset, forecasters could choose a percentage confidence for

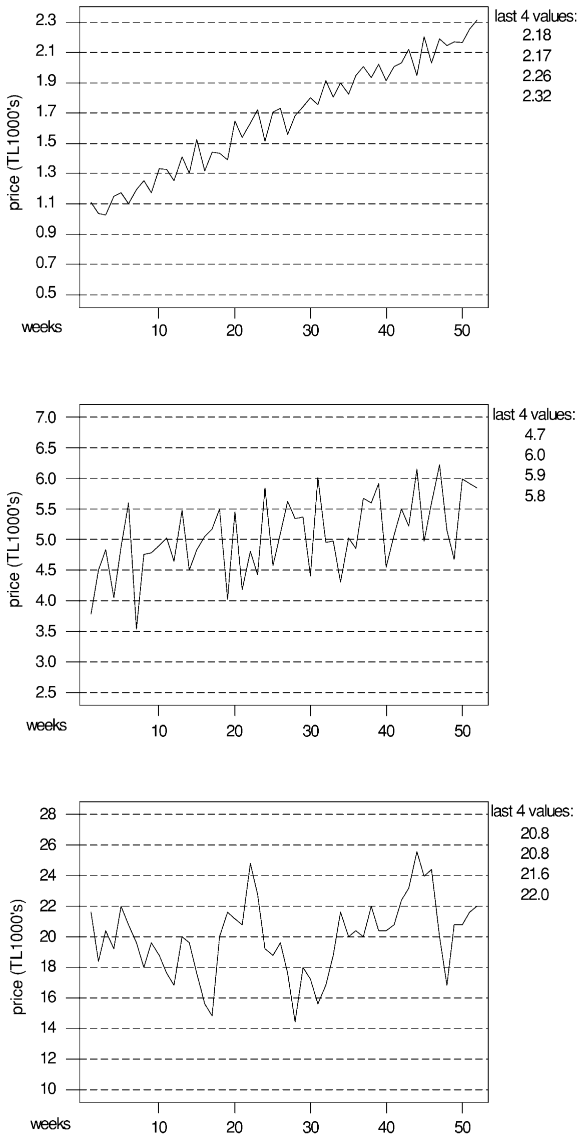

Fig. 1. Example stimulus graphs.

their intervals between 50 and 99% inclusive, but then had to use the same percentage for the entire session.

Session 2**:** Three days later the same participants received feedback about the performance of their earlier forecasts before making a further set of forecasts for 32 different time series. Again forecasters were free to choose their confidence percentage, but again had to employ it for the entire session.

Session 3**:** Three days later the procedure for Session 2 was repeated (i.e. following feedback, forecasters chose their confidence percentage and made predictions for a new set of 32 time series).

Session 4: After another 3 days, feedback from Session 3 was given, followed by the forecasters completing a questionnaire. Finally, the participants were debriefed.

At the start of the first session the participants were given a single sheet of written instructions in Turkish which explained the task and the general nature of the experiment. A specific example was also given of how to construct well-calibrated confidence intervals. An English translation of these instructions is given in Appendix A. The written instructions were reinforced verbally and any questions were answered by the experimenters.

4. Results

4.1. Calibration and overconfidence

> *Hypothesis 1*: Interval forecasts will initially manifest overconfidence but this will be significantly reduced as forecasters receive feedback about their performance.

This hypothesis is strongly supported. The calibration curves for each group show clear improvement over sessions although there is a suggestion that the forecasters may become underconfident with additional practice—the best calibrated group to begin with became clearly underconfident by the third session. The overall pattern for the four groups was similar, however, so the composite curves are shown in Fig. 2 for simplicity. The ANOVA results support the conclusion derived from inspection of the calibration curves as there was a significant main effect of session on calibration ($F_{2,270}$=33.7, P<0.001) with calibration improving between both the first and second, and the second and third sessions.

> *Hypothesis 2*: Improvement in calibration due to feedback will be the result of forecasters being better able to reflect in their confidence intervals the variability in the time series.

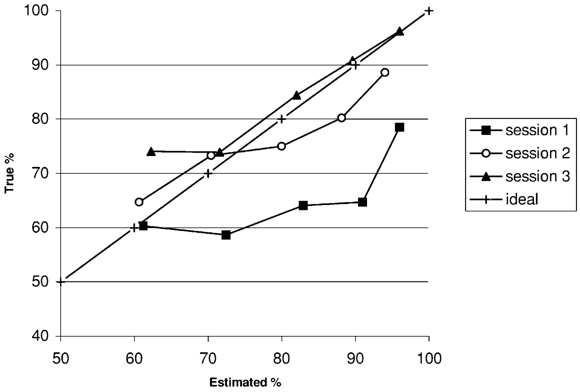

Fig. 2. Calibration curves for all subjects by experimental session.

Taken together, the width of the confidence interval and the percent confidence can be regarded as an estimate of the standard deviation of the stimulus series:

$$\sigma = \frac{x - \text{ц}}{z}$$

where $(x - \text{ц})$ is half the confidence interval width and z is the percent confidence expressed as a z score[2]. By this method we calculated each participant's estimated standard deviation for each stimulus series and correlated these with the true value of the standard deviation in the series so as to produce three correlation coefficients for each participant, one for each of the three sessions. The average correlation between actual and estimated standard deviation for each session was $r =$ 0.85, 0.90 and 0.92, respectively. After transforming these correlation coefficients using a Fisher's z transformation, we performed a completely within-subjects ANOVA which showed a significant improvement in the ability of our forecasters to match their probabilistic responses to the changes in uncertainty in the stimulus series ($F_{2,276} = 63.09, P < 0.001$). Hypothesis 2 was therefore supported.

4.2. Accuracy vs. informativeness

> *Hypothesis 3*: Over sessions there will be a decrease in the width of the judgment intervals provided and an accompanying decrease in the confidence levels selected.

Contrary to our expectations, there was a significant *increase* in the average width of the confidence interval given (1.68 in session 1, 2.33 in session 2 and 2.79 in session 3; $F_{2,276} = 96.27, P < 0.001$). Also contrary to our hypothesis the percent usage of each response category (see Fig. 3: categories 1 to 5 are defined as 50–59%, 60–69%, 70–79%, 80–89% and 90–99%, respectively) shows a slight tendency for the forecasters to shift to giving *higher* confidence percentages over the three sessions (ANOVA showed this to be a statistically significant tendency: $F_{2,270} = 7.43, P = 0.001$). However, a more accurate description of the pattern in Fig. 3 is that responses become more extreme. Further, it is interesting to note that confidence levels actually drop after first set of feedback, before increasing again after the second set of feedback. Overall, it seems that improvement in hit rates, rather than increases in confidence, provide the main thrust in improving the participants' interval calibration. To be specific, hit rate steadily improves over sessions ($F_{2,270} = 105.94, P < 0.001$)—from a mean hit rate of approximately 67% in Session 1 to about 77% in Session 2 to roughly 87% in Session 3. This provides further support for Hypothesis 1 but it seems that Hypothesis 3 is definitely not supported.

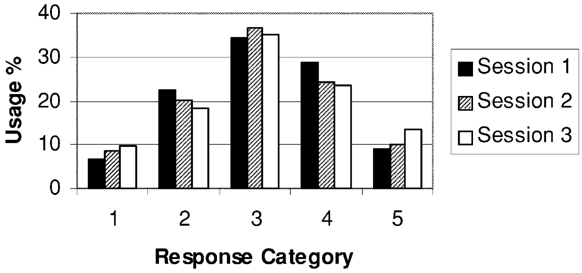

Fig. 3. Distributions of probability usage by experimental session.

[2] This assumes a normally distributed probability distribution function symmetrical around the 'best guess' of the position of the next value in the stimulus series—this assumption may not be correct, particularly for the trended series, but unfortunately we have no way of checking this from our data.

5. Discussion and conclusions

It seems that in a judgmental interval forecasting task people can quite quickly become well-calibrated when provided with calibration feedback. This supports the idea that previous failures to improve calibration through training are the consequence of unfavourable task features. However, an alternative interpretation of our findings is possible. As already mentioned, a persistent finding of research into the calibration of subjective probability judgment is the so-called 'hard–easy effect' whereby overconfidence is observed for poor performance (hit rates below about 70%), underconfidence where performance is good (hit rates above about 80%), and good calibration at levels of performance between these extremes (see, e.g. Lichtenstein et al., 1982; Suantak et al., 1996). In our study, we observed a similar pattern of results including some underconfidence in the final session where the average hit rate was 87%. In fact, the correlation between the average hit rate and under/overconfidence (*n*=12: four groups times three sessions) was $r = -0.88$. It is therefore consistent with our results that the participants achieved good calibration simply by increasing the width of the confidence intervals they gave, thereby improving hit rate, whilst maintaining more-or-less the same level of confidence, or possibly increasing it. If this were the case then we might anticipate increased underconfidence if further practice were given. In other words, our forecasters may not have learnt either to match the width of intervals to the confidence percent given or to match their combined responses (i.e. both the width of the interval and their confidence) to the true level of uncertainty in the series and thus cannot really be said to have improved either their calibration or their discrimination (i.e. their ability to match their uncertainty responses to changes in uncertainty in the stimulus materials) as a result of the feedback given (see, e.g. Yates, 1982).

If such a 'hedging' strategy was widely used, the average confidence should remain approximately constant or increase across the three sessions (which it did: 82, 81 and 84%), whereas there should be a significant increase in the average width of the confidence interval given (which, as we have already seen, was also the case: 1.68, 2.33, 2.79). However, a slightly more fine-grained analysis reveals that the correlation between the width of confidence interval and level of confidence increases over the three sessions ($r = 0.37$, 0.50 and 0.51, respectively) which suggests that our forecasters were at least learning to match confidence-interval width to confidence percentage.

Perhaps more compelling evidence against such a hedging strategy comes from our analysis of improvement in our forecasters' discrimination which we performed in order to test Hypothesis 2. Thus, it appears that *both* the calibration and discrimination of our forecasters improved over the three sessions. It should be noted here that our explicit performance feedback was intended to improve calibration rather than discrimination. According to Stone and Opel (2000), it is only environment feedback that should improve discrimination—as noted earlier, our task contained such feedback implicitly, and therefore it must be to this feedback that we attribute the observed improvements in discrimination.

The improvement in calibration that we observed therefore seems to be the result of at least two tendencies: first, to increase interval width whilst keeping percent confidence constant (thereby increasing hit rate and, initially at least, improving calibration); second, to match probability responses (interval width and percent confidence) better to the true variance in the series. Both these tendencies are compatible with the view of overconfidence (and the hard–easy effect) being the consequence of response error. If it is assumed that confidence intervals are initially too narrow for both the given percent confidence and true variance in the series, then this error is subsequently removed by the application of the two tendencies described above.

An alternative account is that our forecasters initially underestimated the amount of variance in the series but gave appropriate probability responses. Subsequently, their ability to estimate the variance in the series improved which is then manifest in their more appropriate responses. In other words, the initial error may have been in the estimation of the variance rather than in the expression of an accurate estimate as a probability response. We cannot differentiate these two accounts on the basis of our data but there is some evidence that people are quite good at estimating the variance of time series (Fike, 1977; Fike & Ferrell, 1977, 1978), which would seem to support the response-error account, though it is possible that both types of error are present (see, e.g. Juslin, Olsson, & Bjorkman, 1997).

Note that neither account explains why the initial bias occurs in our experiment. In general-knowledge tasks people may assume they will attain an intermediate level of performance (e.g. 75% for two-alternative, forced-choice questions) whereas the questions usually used in studies produce lower levels of performance (see, e.g. Suantak et al., 1996, Fig. 11). It is more difficult to explain the initial bias in this way in our task because it is not clear what an intermediate amount of variance might be, and in any case, the true amount of variance varied across the stimulus series. However, it is possible that the average level of variance in our series was greater than what our participants expected on the basis of their experience with real Turkish stocks or their overall expectations from the stock market at the time the experiment was conducted (even though they were told that the series were not real-time).

From a practical perspective our results suggest that training forecasters in probabilistic forecasting may be a realistic proposition although there is a question mark over the generalizability of any such training to different types of series from those in the training set. Further, as we have already discussed, there is a suggestion from these data that forecasters may 'overshoot' and become underconfident. Thus further experiments with more sessions and different types of series need to be conducted to provide detailed answers to these questions and to permit the design of appropriate training programmes.

The fact that there was no strong tendency to try and attain good calibration by picking the top level of confidence (99%) and giving huge intervals—the ultimate expression of the hedging strategy described earlier—provides some support for Yaniv and Foster's (1995, 1997) assertion that people prefer to give narrower, more informative intervals—it also suggests that participants in our experiment were not simply treating the exercise as a game. Some further comments from the participants explicitly referring to a desire to avoid the use of very wide intervals support these claims:

> 'they would be meaningless'
> 'they would have no value to whoever wants to use them'
> 'how could they possibly help in picking stocks if they're really wide'
> 'it doesn't make sense because it's so easy to give very wide intervals, choose 99% and get perfect match—but what worth could this have for people making investment decisions'

However, the desire to be informative does not seem to us an explanation for overconfidence as it begs the question of why people do not just give lower percent confidence if they prefer narrower intervals. One reason why this tendency to give lower confidence intervals may not have been observed in this study is that, as we suggest in the Introduction, the communicative goals of the task are secondary to achieving good calibration. Thus if we had increased the number of sessions then we may have observed reduction of both confidence interval width and corresponding stated levels of confidence. Another possibility, though, is that our student participants were already used to the convention of 95% confidence intervals and that this acted as an anchor for their chosen confidence levels. If so then this would act as a strong deterrent to reducing confidence levels—the same would apply to most professional forecasters.

Finally, we wish to comment briefly on the similarities and differences between our study and that of O'Connor and Lawrence (1989). As with our study, these authors examined the effects of feedback on judgmental confidence intervals in forecasting from graphical time series. However, as we indicated in the Introduction, the similarities between the two studies are largely superficial, there being substantive differences including the nature of the stimulus series, the kind of judgmental confidence intervals elicited, the type of feedback, and consequently the form of analyses that could be carried out. Despite all these differences, it is interesting to note that our conclusions are much the same: that judgmental confidence intervals were initially overconfident but improved significantly after feedback. That two rather disparate approaches arrive at the same conclusions suggests that these findings are robust, hence highlighting the potential for using prediction intervals for reliably communicating uncertainties in diverse domains.

Acknowledgements

The authors wish to acknowledge the assistance of Amir Aghaty, Eda Serdaroglu, and Emrah Ayaydin in the collection of the data. We would also like to thank Russ Ferrell, Alastair McClelland, Marcus O'Connor and some anonymous referees for their comments on earlier drafts of the paper.

Appendix A

Dear Participant:

In this study, we request that you examine the 32 time series presented and make forecasts for each series. The time series presented to you show the weekly closing prices of various stocks taken from various time periods in previous years. We would like you to convey your one-period-ahead predictions via *interval forecasts* for the next week's (i.e. next Friday's) closing price. To do so, you first need to choose a confidence percentage (between 50 and 99%) that you feel comfortable with. *Please note that you are required to use the same confidence percentage for all the stocks (i.e. time series) within the same session. If you would like to or if you think you need to, you may change your percentage in other sessions*. After specifying your confidence percentage, you are asked to specify the lowest and the highest possible values (with your selected confidence percentage) that a particular stock's closing price could assume for the next period. For example, let's say you've chosen 75% as your confidence percentage[3]. This means that, given 100 prediction intervals you've specified, you expect 75 of them to include the realized value, while expecting the realized values to fall outside of the limits you've given for the remaining 25 intervals.

Our study consists of four sessions. In the first, second and third sessions, you'll be asked to construct your interval forecasts. Accordingly, you'll be receiving feedback in the beginning of second, third and fourth sessions. You will not be asked to make forecasts in the fourth session.

THANK YOU FOR YOUR PARTICIPATION IN THIS STUDY.

[3] Half the participants received 75% as the figure in the example whilst the other half received 90%. Our analyses showed a significant anchoring effect due to these different examples such that participants tended to choose a higher confidence value if presented with the 90% example than the 75%. There were no other discernible effects of these different anchors on calibration, however.

References

Adams, P. A., & Adams, J. K. (1958). Training in confidence judgments. *American Journal of Psychology*, *71*, 747–751.

Alpert, M., & Raiffa, H. (1982). A progress report on the training of probability assessors. In Kahneman, D., et al. (Ed.), *Judgment under uncertainty: heuristics and biases*. Cambridge: Cambridge University Press, pp. 294–305.

Arkes, H. R. (2001). Overconfidence in judgmental forecasting. In Armstrong, J. S. (Ed.), *Principles of forecasting: a handbook for researchers and practitioners*. Boston: Kluwer Academic Publishers, pp. 495–515.

Ayton, P., & McClelland, A. G. R. (1997). How real is overconfidence? *Journal of Behavioral Decision Making*, *10*, 279–285.

Balzer, W. K., Doherty, M. E., & O'Connor Jr., R. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin*, *106*, 410–433.

Balzer, W. K., Hammer, L. B., Sumner, K. E., Birchenough, T. R., Martens, S. P., & Raymark, P. H. (1994). Effects of cognitive feedback components, display format and elaboration on performance. *Organizational Behavior and Human Decision Processes*, *58*, 369–385.

Balzer, W. K., Sulsky, L. M., Hammer, L. B., & Sumner, K. E. (1992). Task information, cognitive information, or functional validity information: which components of cognitive feedback affect performance? *Organizational Behavior and Human Decision Processes*, *53*, 35–54.

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

Bornstein, B. H., & Zickafoose, D. J. (1999). 'I know I know it, I know I saw it': the stability of the confidence–accuracy relationship across domains. *Journal of Experimental Psychology: Applied*, *5*, 76–88.

Fike, R. L. (1977). *Subjective Estimation of Variability*. Unpublished doctoral dissertation, The University of Arizona.

Fike, R. L., & Ferrell, W. R. (1978). Subjective detection of differences in variance from small samples. *Organizational Behavior and Human Performance*, *22*, 262–278.

Fike, R. L., & Ferrell, W. R. (1977). Heuristics in variability estimation. *In Proceedings of the 1977 International Conference on Cybernetics and Society*, Washington, DC.

Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence. *Psychological Review*, *98*, 506–528.

Juslin, P., Olsson, H., & Bjorkman, M. (1997). Brunswikian and Thurstonian origins of bias in probability assessment: on the interpretation of stochastic components of judgment. *Journal of Behavioral Decision Making*, *10*, 189–209.

Keren, G. (1987). Facing uncertainty in the game of bridge: a calibration study. *Organizational Behavior and Human Decision Processes*, *39*, 98–114.

Keren, G. (1991). Calibration and probability judgments: conceptual and methodological issues. *Acta Psychologica*, *77*, 217–273.

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance*, *26*, 149–171.

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

Makridakis, S., Andersen, A., Carbone, E., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., & Winkler, R. (1982). The accuracy of extrapolation (time series) methods: results of a forecasting competition. *Journal of Forecasting*, *1*, 111–153.

Murphy, A. H., & Winkler, R. L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association*, *79*, 489–500.

O'Connor, M., & Lawrence, M. (1989). An examination of the accuracy of judgmental confidence intervals in time series forecasting. *Journal of Forecasting*, *8*, 141–155.

Oskamp, S. (1962). The relationship of clinical experience and training methods to several criteria of clinical prediction. *Psychological Monographs*, *76*, 1–25.

Önkal, D., & Muradoğlu, G. (1995). Effects of feedback on probabilistic forecasts of stock prices. *International Journal of Forecasting*, *11*, 307–319.

Önkal-Atay, D., Thomson, M. E., & Pollock, A. C. (2002). Judgemental forecasting. In Clements, M. P., & Hendry, D. F. (Eds.), *A companion to economic forecasting*. Oxford: Blackwell, pp. 133–151.

Plous, S. (1995). A comparison of strategies for reducing interval overconfidence in group judgments. *Journal of Applied Psychology*, *80*, 443–454.

Powers, W.T. (1973). Feedback: beyond behaviorism. *Science*, *179*, 351–356.

Remus, W., O'Connor, M., & Griggs, K. (1996). Does feedback improve the accuracy of recurrent judgmental forecasts? *Organizational Behavior and Human Decision Processes*, *66*, 22–30.

Roth, P. L. (1993). Research trends in judgment and their implications for the Schmidt–Hunter global estimation procedure. *Organizational Behavior and Human Decision Processes*, *54*, 299–319.

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

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*, 271–283.

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*, 282–309.

Suantak, L., Bolger, F., & Ferrell, W. R. (1996). The hard–easy effect in subjective probability calibration. *Organizational Behavior and Human Decision Processes*, *67*, 201–221.

Todd, F. J., & Hammond, K. R. (1965). Differential effects in two multiple-cue probability learning tasks. *Behavioral Science*, *10*, 429–435.

Weiner, N. (1948). *Cybernetics: control and communication in the animal and the machine*. New York: Wiley.

Wright, G., & Ayton, P. (1986). Subjective confidence in forecasts: a response to Fischhoff and MacGregor. *Journal of Forecasting*, *5*, 117–123.

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

Yaniv, I., & Foster, D. P. (1997). Precision and accuracy of judgmental estimation. *Journal of Behavioral Decision Making*, *10*, 21–32.

Yates, J. F. (1982). External correspondence: decompositions of the mean probability score. *Organizational Behavior and Human Performance*, *30*, 132–156.

Biographies: Fergus BOLGER, during the preparation of this paper, was Associate Professor of Management at Bilkent University, Ankara, Turkey. He is now a Lecturer in Psychology at the University of Leicester, England. He holds a PhD. in Cognitive Psychology from the University of London, England.

Dilek ÖNKAL-ATAY is Associate Professor of Decision Sciences at Bilkent University, Ankara, Turkey. She holds a PhD. in Decision Sciences from the University of Minnesota, USA.

Both authors have research interests in subjective probability judgment and judgmental forecasting in particular, and judgment and decision making in general. They have published in *International Journal of Forecasting, Journal of Behavioral Decision Making, Journal of Forecasting,* and *Organizational Behavior and Human Decision Processes* amongst others.