



ELSEVIER

International Journal of Forecasting 11 (1995) 307–319

international journal  
of forecasting

# Effects of feedback on probabilistic forecasts of stock prices

Dilek Önkal\*, Gulnur Muradoğlu

*Faculty of Business Administration, Bilkent University, 06533 Ankara, Turkey*

---

## Abstract

This paper reports the results of an experiment in stock-price forecasting that investigated the effects of feedback on various dimensions of probability forecasting accuracy. Three types of feedback were used: (1) simple outcome feedback, (2) outcome feedback presented in the task format, and (3) performance feedback in the form of an overall accuracy score in addition to detailed calibration information. While calibration improved for all the feedback groups, forecasters' skill was found to improve only for the task-formated outcome feedback and performance feedback groups (but not for the simple outcome feedback group). Finally, the forecasters in the performance feedback group also improved their mean slope and mean probability scores, an effect not observed in the other feedback groups. It is suggested that, in a dynamic environment like the stock market, probability forecasting offers distinct advantages by providing an important channel of communication between the forecasters and the users of financial information.

*Keywords:* Probability forecasting; Judgmental forecasting; Stock-price forecasting; Outcome feedback; Performance feedback; Calibration

---

## 1. Introduction

This study aims to explore the effects of different types of feedback on the quality of judgmental probability forecasts given by individual forecasters. In particular, we focus on the use of outcome feedback and performance feedback within the framework of probabilistic forecasting of stock prices. Outcome feedback can be defined as information about the realization of a previously predicted event. Performance feedback gives information about the accuracy of the forecaster's predictions, based on both the fore-

caster's predictions and the outcomes that actually occur.

The crucial role played by feedback in probability assessment tasks has been accentuated repeatedly in the literature. In his review of the literature on subjective probability assessments and related cognitive processes, Hogarth (1975, p.278) concluded that "...substantive experts can make meaningful assessments in situations where they make forecasts over a period of trials and receive feedback as to the accuracy of their judgments". Also, the provision of feedback regarding the correspondence of forecasts with actual occurrences was viewed by Moriarty (1985) as a critical design feature of forecasting systems that involve management judgment.

---

\* Corresponding author.

Given this overall emphasis on feedback in forecasting settings, it is surprising that there have only been a limited number of empirical studies examining this issue. Reporting the results of an empirical study on outcome feedback, Fischer (1982) has suggested that the mere knowledge of results (i.e. outcome feedback) is ineffective in improving the overall accuracy of probability forecasts. Regarding the effectiveness of performance feedback, previous studies have dealt with (1) scoring-rule feedback and (2) calibration feedback.

A scoring rule assigns an overall score to a forecaster, based on a function of the forecaster's reported probability forecasts and the outcomes that actually occur, computed over a set of probability forecasts (Winkler, 1969, and Friedman, 1983). Previous empirical studies have revealed mixed results with respect to the effectiveness of scoring-rule feedback on the accuracy of forecasts. Stael von Holstein (1972) and Fischer (1982) have concluded that the provision of scores from such rules have no effect on the performances of their forecasters. In contrast, Kidd (1973) (cited in Beach, 1975) has shown that scoring-rule feedback could be effective in improving forecasters' accuracy levels.

Calibration feedback involves giving forecasters information about their ability to assign appropriate probabilities to future outcomes. If for all the predicted outcomes that a forecaster assigns a given probability, the proportion of those outcomes that actually occur is equal to the assigned probability, then that forecaster is said to be well-calibrated. Utilizing mainly general knowledge tasks, numerous empirical studies have shown some improvement in calibration as a result of feedback (Adams and Adams, 1958, 1961; Oskamp, 1962; Schaefer and Borcharding, 1973; Pickhardt and Wallace, 1974; and Lichtenstein and Fischhoff, 1980). Comparatively few studies, however, have been conducted within the probability forecasting domain. Murphy and Daan (1984) and Murphy et al. (1985) have found calibration feedback to be effective in field studies of weather forecasters. Benson and Önkal (1992) have reported the results of a

laboratory experiment where they found the provision of calibration feedback to improve the performance of probability forecasters. Additional studies in the forecasting domain are needed—in part, because there has been some controversy regarding the generalizability of results from general-knowledge tasks to forecasting tasks. In particular, while Fischhoff and MacGregor (1982) have argued that the results from studies using almanac questions are generalizable, Wright and Ayton (1986) and Ronis and Yates (1987) have disputed these arguments.

In this paper we examine the effects of outcome and performance feedback on various accuracy dimensions of probabilistic forecasts regarding stock prices. We adopt the experimental framework of Yates et al. (1991). They investigated probabilistic forecast accuracy for stock prices and earnings. The authors aimed (1) to re-examine previous results on the accuracy of probability judgments on securities, and (2) to test the existence of an inverse relationship between expertise and accuracy. They found the probabilistic forecasts of changes in stock prices to be quite inaccurate. The authors also observed an inverse expertise effect between graduate and undergraduate students, with undergraduates being more accurate than graduates.

We adapt the procedure employed by Yates et al. (1991) to the Turkish stock market and extend their work to delineate the effects of feedback on probabilistic forecasts of stock prices. We utilize three types of feedback: (1) simple outcome feedback (i.e. mere knowledge of results), (2) outcome feedback presented in the same format as the structure of the task confronting forecasters, and (3) performance feedback in the form of calibration feedback along with an overall accuracy score.

## **2. The setting: An emerging securities market**

Attempts to liberalize the highly inefficient and strictly regulated financial markets in Turkey

started at the beginning of the 1980s. Although the establishment of the legal framework and regulatory agencies for the stock market were completed in 1982, the Istanbul Securities Exchange (the only stock exchange in Turkey) was established in 1986. Until 1987, the employees of the stock exchange could hold stock portfolios without notification. It was in 1990 that legislation against insider trading was passed for the first time. At the end of 1991, when this study was conducted, 123 stocks were traded at the Istanbul Securities Exchange, and the volume of trade amounted to approximately \$650 million. Out of the 162 intermediaries and brokerage houses, 60 were bank-affiliated companies.

Stock price forecasts in the United States have been shown to be relatively inaccurate when compared with earnings forecasts (Yates et al., 1991). This can be attributed to the efficiency of the stock market in the United States. If the market is efficient, all relevant information including knowledge of previous prices (Fama, 1965), public announcements (Ball and Brown, 1968), and even monopolistic information (Jensen, 1968) is fully reflected by the stock prices, so that no investor can beat the market continuously. In efficient markets, therefore, no particular investment method is assumed to be superior to the random selection of investment portfolios.

Predicting stock prices in the Istanbul Securities Exchange could be viewed as being easier than predicting stock prices in a developed market (e.g. the New York Stock Exchange), due to the inefficiency of the market. The Istanbul Securities Exchange is known to be weak-form (Muradoglu and Oktay, 1993, and Muradoglu and Unal, 1993) and semi-strong form (Muradoglu and Önkal, 1992) inefficient. This means that prudent investors can adopt certain trading rules based on stock-price forecasts and earn above-normal profits. Furthermore, the relatively small number of stocks in the Istanbul Securities Exchange in comparison with the ones in developed countries decreases the complexity for the investor. Hence, in the Istanbul stock market, there may be a potential for improving stock-price forecasting perform-

ance. We attempt to determine if various forms of feedback can achieve this potential.

### 3. The method

Subjects for the study were recruited from the graduate and undergraduate classes of the Faculty of Business Administration of Bilkent University. The purpose of the study was described in preposted announcements. No monetary nor non-monetary bonuses were offered, aside from: (1) the opportunity to evaluate possible investment alternatives in a real stock market setting, and (2) the opportunity to evaluate and improve probabilistic forecasting skills.

Sixty-eight subjects completed the four-week-long experiment. The subjects were randomly assigned to three feedback groups: (1) simple outcome feedback, (2) task-formated outcome feedback, and (3) performance feedback. Feedback groups were comprised of 19, 24, and 25 subjects, respectively.

The experiment involved four weekly forecasting sessions, and the task was to provide probability forecasts of the closing stock prices of 34 companies listed in the Istanbul Stock Exchange. The choice of the companies was made deliberately to minimize task complexity. Accordingly, a list of 34 stocks with the highest volume of trade during the preceding 52-week period was selected. For each stock, the subjects were asked to make forecasts regarding the weekly price change, i.e. the percentage change in the closing price of stock between the previous Friday and the current Friday. Subjects provided their forecasts in the form of subjective probabilities that conveyed their degrees of belief. In particular, the subjects were asked to complete a response form for each company (Table 1).

The ranges of the stock-price change in the response form were constructed by considering the weekly price changes of the stock market index during the previous 52-week period. During the previous 52 weeks, the average weekly price change was 3%, with the maximum increase being 8% and the maximum decrease

Table 1

	Weekly price change interval (in percentages, Friday to Friday)	Probability
(6)	Increase more than 10%	—%
(5)	Increase 5–10%	—%
(4)	Increase up to 5%	—%
(3)	Decrease 0–5%	—%
(2)	Decrease 5–10%	—%
(1)	Decrease more than 10%	—%
		—%

being 5%. The first 5% increase range was designed to contain the average increase, the second to contain the maximum increase during the previous year, while the third range was designed for highly volatile stocks. Intervals (1)–(3) were designed to be symmetric to intervals (4)–(6) for cognitive purposes. It should be emphasized that intervals (1) and (6) were included to accentuate the high volatility that could easily be observed in emerging markets with expanding volumes of trade.

At the beginning of the first session, all subjects were given detailed information about the design and goals of the study. Definitions of 'subjective probability' and 'probability forecasting tasks' were provided with illustrative examples. Subjects in the performance feedback group were given additional information about the performance measures used in the experiment, along with specific examples. Subjects were informed that they could achieve the best possible scores only by expressing true opinions, thereby avoiding hedging.

At the beginning of the first session, subjects were presented with folders containing three separate forms: background forms, response sheets, and questionnaires. The background forms contained information regarding the name of each company, its industry, its net profits as of the end of the third quarter of 1991, earnings per share, and the price–earnings ratio as of the last day of the preceding week. The folders also provided the weekly closing stock prices (i.e. the closing stock prices for each Friday) of the preceding 3 months (12 weeks) in tabular form,

as well as the weekly closing prices for the last 52 weeks in graphical form. Response sheets were comprised of the response forms (Table 1) and instructions about the forecasting task. The questionnaire was completed after the end of the fourth session. This instrument was designed to provide information about the participants' field of study, year in school, names of previous and current finance and decision analysis courses, previous and current experience in the stock market and its duration, and previous and current experience in trading and its duration. Subjects were also asked to delineate the sources they utilized in making their forecasts and to provide a ranking with respect to the frequency of usage.

In order to duplicate real forecasting settings, the subjects were allowed to take the background folders home. They were given the experimental material on Monday morning and were requested to submit the completed response sheets by Tuesday morning at 9 a.m., before the opening of the session at the stock exchange. They were to complete their forecasts by the next day, since each additional day would give them more information about the current week's closing stock prices. They were also permitted to utilize any information source they preferred in making their forecasts, excluding the other participants of the study.

At the beginning of the second, third, and fourth sessions, feedback from all previous sessions were made available to the subjects. Each feedback group met at a different session and received individual attention. Also, subjects were encouraged to discuss their personal feedback with the experimenters. Details of the feedback given to each of the three groups are presented in the following subsections.

### 3.1. Simple outcome feedback group

The simple outcome feedback group served as a control group for the experiment. As feedback, subjects in this group were given the previous Friday's closing price marked on the graphical and tabular information forms provided in the background folders for each of the 34 com-

panies. Although these prices were available to all subjects through the media, this prepared format for the simple outcome feedback group was chosen so as not to increase this group's perceived task difficulty in comparison with the other feedback groups.

### 3.2. Task-formated outcome feedback group

Task-formated outcome feedback included (1) feedback given to the simple outcome feedback group subjects, and (2) the realized weekly price change for each of the stocks, marked on each subject's response sheets adjacent to their probability assessments from the previous week.

### 3.3. Performance feedback group

Subjects in the performance feedback group received (1) feedback given to the task-formated outcome feedback group subjects, (2) their individual mean probability scores (as defined below) computed from their forecasts of the previous week, (3) their individual calibration scores (as defined below) computed from the previous week's forecasts, and (4) their overall mean probability forecasts and their overall mean proportion of correct forecasts for each of the price change intervals.

The mean probability score for multiple events, PSM (Yates, 1988) was used to evaluate the overall performance of subjects. (See the appendix for the definition of this criterion.) The mean probability score has partitioning properties that facilitate the identification of various aspects of forecaster performance (Murphy, 1972a,b, 1973, and Yates, 1988).

Calibration is the most widely used performance criterion to emanate from these decompositions (Lichtenstein et al., 1982). Calibration provides information about the forecaster's ability to assign appropriate probabilities to outcomes. A forecaster is well-calibrated if, for all predicted outcomes assigned a given probability, the proportion of those outcomes that occur (i.e. the proportion correct) is equal to the probability that is assigned by the subject. (See the appendix for the definition of this criterion.) Subjects in

the performance feedback group were given their mean probability forecasts and proportion correct for each of the six predefined weekly price change intervals besides their calibration scores in order (1) to help their understanding of the calibration scores attained, and (2) to aid their identification of any systematic under- or overforecasting they might be engaged in.

## 4. Findings

Performance measures used to explore the effects of different types of feedback on probabilistic forecasts of stock prices were the mean probability score, calibration score, mean slope, scatter score, forecast profile variance, and the skill score (see the appendix). Using these measures, across-session performances were tracked within each performance group (1) by using the Wilcoxon signed-ranks test to detect the changes in each of the six performance measures session by session, and (2) by providing as standards of comparison the scores that would be attained by a uniform forecaster, a historical forecaster, and a base-rate clairvoyant (Yates et al., 1991).

A 'uniform forecaster' gives 'equally likely' assessments to all the possible outcomes. Since our task involves six intervals, a uniform forecaster would give probability forecasts of 1/6 to all the stocks in question. A 'historical forecaster' gives forecasts identical to the historical relative frequencies. Given the volatility of the stock market under consideration, we set the historical forecaster's probability forecasts equal to the relative frequencies realized in the previous week. A 'base-rate clairvoyant' is a forecaster who can perfectly foretell the relative frequencies (i.e. base rates) with which the price changes will occur.

Table 2 presents the medians for the six performance measures attained by each experimental group during each session. Also provided are the scores that would be attained by a uniform forecaster, a historical forecaster, and a base-rate clairvoyant. Statistically significant changes ( $p$ -values less than 0.10) from one session to the next and from the first session to

Table 2

Median values for various performance measures for the simple outcome feedback group (SOF), task-formated outcome feedback group (TOF), and performance feedback group (PF), with corresponding measures for uniform (U) and historical (H) forecasters, and for base-rate clairvoyants (B)

Performance measure		Session			
		1	2	3	4
PSM	SOF	0.873	1.228*** <sup>W</sup>	0.916*** <sup>B</sup>	0.882
	TOF	0.914	1.229*** <sup>W</sup>	0.889*** <sup>B</sup>	0.891
	PF	0.887	1.194*** <sup>W</sup>	0.907*** <sup>B</sup>	0.818*** <sup>B**L</sup>
	U	0.833	0.833	0.833	0.833
	H	0.778	1.066	1.102	0.811
	B	0.754	0.706	0.725	0.777
Calibration	SOF	0.061	0.369*** <sup>W</sup>	0.106*** <sup>B</sup>	0.035*** <sup>B**L</sup>
	TOF	0.088	0.373*** <sup>W</sup>	0.076*** <sup>B</sup>	0.030*** <sup>B**L</sup>
	PF	0.078	0.365*** <sup>W</sup>	0.089*** <sup>B</sup>	0.029*** <sup>B***L</sup>
	U	0.079	0.127	0.108	0.057
	H	0.024	0.360	0.377	0.035
	B	0.000	0.000	0.000	0.000
Mean slope	SOF	0.022	-0.031*** <sup>W</sup>	-0.025	0.023*** <sup>B</sup>
	TOF	0.020	-0.027*** <sup>W</sup>	-0.015	0.012*** <sup>B</sup>
	PF	0.016	-0.021*** <sup>W</sup>	-0.028	0.032*** <sup>B***L</sup>
	U	0.000	0.000	0.000	0.000
	H	0.000	0.000	0.000	0.000
	B	0.000	0.000	0.000	0.000
Scatter	SOF	0.074	0.142	0.089* <sup>B</sup>	0.111
	TOF	0.079	0.134	0.087* <sup>B</sup>	0.081
	PF	0.085	0.097	0.070	0.058
	U	0.000	0.000	0.000	0.000
	H	0.000	0.000	0.000	0.000
	B	0.000	0.000	0.000	0.000
F. profile variance	SOF	0.038	0.039	0.029	0.034
	TOF	0.034	0.038	0.028** <sup>B</sup>	0.033
	PF	0.034	0.032	0.025	0.030
	U	0.000	0.000	0.000	0.000
	H	0.023	0.013	0.021	0.018
	B	0.013	0.021	0.018	0.009
Skill	SOF	0.120	0.522*** <sup>W</sup>	0.191*** <sup>B</sup>	0.106*** <sup>B</sup>
	TOF	0.159	0.523*** <sup>W</sup>	0.164*** <sup>B</sup>	0.114*** <sup>B**L</sup>
	PF	0.132	0.488*** <sup>W</sup>	0.182*** <sup>B</sup>	0.041*** <sup>B***L</sup>
	U	0.079	0.127	0.108	0.057
	H	0.024	0.360	0.377	0.036
	B	0.000	0.000	0.000	0.000
Session volatility		0.755	0.706	0.725	0.777

\*  $p$ -value < 0.10.

\*\*  $p$ -value < 0.05.

\*\*\*  $p$ -value < 0.001.

(Wilcoxon signed-ranks test used.)

<sup>F</sup> First session better than last session.

<sup>L</sup> Last session better than first session.

<sup>W</sup> Worse than previous section.

<sup>B</sup> Better than previous session.

the last are denoted by superscripts defined in the footnote to the table.

#### 4.1. Simple outcome feedback group

The simple outcome feedback group, which received realized prices as the only feedback, showed improved calibration scores (i.e. the last session's calibration performance was better than that of the first session,  $p = 0.097$ ) with no apparent deteriorations in other scores. This signifies that, even with simple outcome feedback and continuous training, investors can improve their ability to assign probabilities that match the actual relative frequencies of future outcomes.

This group's median calibration score was better (i.e. lower) than that of the uniform forecaster, the same as the historical forecaster, and worse than the base-rate clairvoyant by the end of the fourth session. A calibration score better than that of a uniform forecaster suggests that the probability assessments of simple outcome feedback group subjects show a better overall performance in comparison with total uncertainty/ignorance. In particular, 68% of these subjects attained calibration scores superior to that of the uniform forecaster in Session 4.

It is also interesting to note that 0% of the simple outcome feedback group subjects obtained calibration scores better than those of the historical forecaster in Session 1. However, 58% of these subjects achieved better calibration score than the historical forecaster in Session 4. The fact that a calibration performance better than that of a historical forecaster was attained has an important policy implication. Instead of employing the widely used historical data, the application of subjective probabilities in investment decisions may result in higher profit opportunities. If feedback improves forecasting abilities, as tentatively suggested by this study, this implies that experts might be trained in using subjective probabilities for better investment decisions.

#### 4.2. Task-formated outcome feedback group

Task-formated outcome feedback group subjects improved their skill scores (the last session's skill score was better than that of the first session with  $p = 0.074$ ) and calibration scores (corresponding  $p = 0.074$ ). It appears that receiving feedback in the task format helped subjects improve their overall forecasting quality, as displayed by their skill scores. This was achieved in addition to the improvement in their proficiency in calibration (as was the only finding with the simple outcome feedback group).

Better skill scores reveal that the task-formated outcome feedback group was able to improve that part of overall accuracy (measured by PSM) that is under the forecaster's control. The fact that the improved PSM score of this group (0.891 in Session 4 as opposed to 0.914 in Session 1) is not statistically significant is mainly due to the volatility of actual outcomes [i.e.  $\text{Var}(d_k)$ ], which are determined by what happens in the forecasting environment. Specifically, the new government's first explicit support for the stock market was announced in week 4, which, in turn, had a pronounced impact on volatility (as can be observed from Table 2).

Similar to the calibration performance observed for the simple outcome feedback group, the task-formated outcome feedback group subjects attained better scores than the uniform and historical forecasters by Session 4. In Session 1, only 38% of these subjects had calibration scores better than the uniform forecaster; whereas in Session 4, 54% of the subjects demonstrated a comparatively superior calibration performance. Also, while 4% of the subjects were better calibrated than the historical forecaster in the first session, 54% were found to show better calibration in the last session. Similar results were found with the skill scores. In Session 1, 13% of the subjects obtained better skill scores than the uniform forecaster, while no subjects surpassed the skill performance of the historical forecaster. However, in Session 4, 25% of the subjects attained better skill scores than the uniform forecaster, and 21% attained scores

better than the historical forecaster. As in the previous group, the base-rate clairvoyant's calibration and skill scores were not surpassed by any of the subjects.

#### 4.3. Performance feedback group

As was expected, provision of performance feedback in the form of PSM and calibration scores, as well as the mean probability forecasts and proportion correct for each of the intervals, resulted in significantly improved forecasting performance over and above the other feedback groups. In addition to the progressions in calibration and skill scores that were observed in the task-formated outcome feedback group, the performance feedback group improved their PSM and mean slope scores significantly ( $p$ -values are 0.0003 for calibration, 0.002 for skill, 0.045 for PSM, and 0.045 for mean slope). Subjects in this group achieved superior accuracy (i.e. lower PSM) mainly through their improved mean slopes. This indicates that the performance feedback group subjects improved their ability to discriminate between occasions when the actual price change did and did not fall into the specified intervals.

In the final session, 80% of the subjects obtained better calibration scores than those of the uniform forecaster, and 60% obtained better scores than those of the historical forecaster. With respect to PSM and skill components, 12% of the subjects obtained better scores (in both PSM and skill) than the uniform forecaster in Session 1, while no subjects had better scores than either the historical forecaster or the base-rate clairvoyant. In Session 4, however, 56% of the subjects had better scores (in PSM and skill) than the uniform forecaster, 40% were better than the historical forecaster, and 16% were actually better than the base-rate clairvoyant. This superior performance was reflected in the mean slope scores, with 88% of the subjects performing better than either the uniform forecaster, the historical forecaster, or the base-rate clairvoyant in the final session. Better scores in comparison with the uniform forecaster reveal that feedback improves forecasting skills in com-

parison with the case of total uncertainty, where the investor would assign equal probabilities to all intervals. Scores better than those of the historical forecaster suggest that the use of probabilistic forecasting can result in better investment decisions than the use of historical data alone.

Multiple-group comparisons of results for each of the four sessions were made via Kruskal-Wallis tests. No significant differences in any of the performance measures were found among the three groups for the first three sessions (all  $p > 0.10$ ). However, in the fourth session, the three groups showed significant differences in PSM ( $p = 0.043$ ), mean slope ( $p = 0.048$ ), and skill ( $p = 0.039$ ). These results are consistent with the findings from within-group comparisons across sessions. As presented above, all three groups improved their calibration by the fourth session, leading to no significant differences in calibration among these groups in Session 4. In addition to enhanced calibration, however, the task-formated outcome feedback group also attained improved skill, while the performance feedback group attained improved skill, mean slope, and PSM. These differential improvements of the three groups led to the significant differences in skill, mean slope, and PSM scores found in the final session.

The potential impact of subjects' use of various external information sources on their performance was examined and no significant correlations were found. The effects of external factors on performance were also investigated by analyzing the correlations between the six performance measures and the following variables: length of active trading experience, number of finance courses taken, and previous exposure to subjective probability concepts (via courses in decision analysis). The only external factor that was found to have a significant relation to performance was the length of financial experience, i.e. active trading. As the length of active trading experience increased, forecast profile variance increased ( $p = 0.042$ ). This may be seen as suggesting that, as subjects gained active trading experience, their forecasts deviated more from those of the uniform fore-



Table 3

Relative frequency distributions of responses in the center ( $\pm 5\%$  price change) vs. the tails (more than 5% price change in either direction) for the simple outcome feedback (SOF), task-formated outcome feedback (TOF), and performance feedback (PF) groups

Probability responses	SOF		TOF		PF	
	Session 1	Session 4	Session 1	Session 4	Session 1	Session 4
<i>Center</i>						
0.00–0.20	0.258	0.116	0.209	0.115	0.269	0.139
0.21–0.40	0.269	0.145	0.228	0.165	0.272	0.176
0.41–0.60	0.296	0.311	0.252	0.401	0.273	0.380
0.61–0.80	0.111	0.193	0.197	0.160	0.127	0.187
0.81–1.00	0.066	0.235	0.114	0.159	0.059	0.118
<i>Tails</i>						
0.00–0.20	0.100	0.303	0.187	0.213	0.115	0.192
0.21–0.40	0.150	0.237	0.219	0.272	0.165	0.281
0.41–0.60	0.309	0.274	0.261	0.338	0.313	0.308
0.61–0.80	0.263	0.115	0.178	0.099	0.240	0.135
0.81–1.00	0.178	0.071	0.155	0.078	0.167	0.084

caster, who assigns equal probabilities to all intervals.

Table 3 displays the relative frequency distributions of subjects' Session 1 and Session 4 responses for the center [i.e. sum of probability forecasts given for intervals (3) and (4) in the response form] and the tails [sum of probability forecasts given for intervals (1), (2), (5), and (6) in the response form] for the three groups. As reflected by their probability usages, feedback in all forms led the subjects to stop putting weight out in the tails. In particular, the usage of all tail probabilities exceeding 0.60 decreased significantly in Session 4, as compared with Session 1 for all feedback groups. Accordingly, the usage of higher probabilities in the center increased for all groups. These findings further confirm Fischer's (1982) result that feedback has the effect of deterring the inappropriate use of large probabilities in the tails.

## 5. Conclusion

This study tested the effects of different types of feedback on the accuracy of stock-price forecasts. The three types of feedback utilized were: (1) simple outcome feedback, (2) outcome feedback provided in the original task format used by

the forecasters, and (3) performance feedback consisting of calibration feedback in addition to an overall accuracy score. Confirming the results of the few previous studies (Murphy and Daan, 1984; Murphy et al., 1985; and Benson and Önkal, 1992), information provided by performance feedback has been shown to improve forecast accuracy. Outcome feedback presented in original task format (which was not employed by previous researchers) improved forecast accuracy in comparison with simple outcome feedback but not in comparison with performance feedback.

We found that feedback in all forms improved the forecasters' ability to assign accurate probabilities to future outcomes that would match actual relative frequencies (i.e. that it improved forecasters' calibration). For all feedback groups, calibration scores were better than those of a uniform forecaster and better than or equal to those of a historical forecaster. Such improvement is important in that it shows that feedback, in any form, can be used to improve probabilistic assessments of future stock prices by investors or analysts.

Outcome feedback presented in original task format and performance feedback improved the skill scores of the subjects: the part of overall accuracy that is under forecasters' control. Having received performance feedback, the subjects

could assess probabilities better than could the uniform forecaster.

Performance feedback also improved the overall accuracy of the forecasters. This effect mainly stemmed from the forecasters' ability to better discriminate between occasions when the actual price change would and would not fall into the specified intervals. The discrimination performance of this group was better than that of the uniform and historical forecasters and the base-rate clairvoyant, while their overall accuracy was better than that of the uniform forecaster (i.e. total uncertainty or ignorance).

It should be noted that the consistent pattern observed in all the feedback groups may partially be attributed to the fluctuations of the market. It may be argued that an emerging market is relatively more volatile than a developed market. Several measures were taken to compensate for such factors and to prevent the special properties of the market from affecting the generality of results. First, a one-week forecast horizon was chosen to ensure that the forecast-period volatility of this study is comparable with the forecast-period volatility of studies conducted in developed markets. Second, the particular stocks employed were selected on the basis of highest volume of trade to reduce volatility. Finally, price change intervals were made comparable with the price change intervals used in other studies. That is, the goal was to match the volatility in those markets via adjustments in the forecast horizon. Future research that would both compensate for market volatility and further reduce task complexity could involve running similar experiments for more iterations using different forecast horizons.

The importance of this study rests on two major findings. First, feedback, independent of its form, improves the ability of forecasters to assign meaningful probabilities to future outcomes in a financial setting. This improvement results in predictions that are better calibrated than those that could be made using historical data. Financial forecasts are frequently reported as point estimates or as forecasts of ranges that do not reveal how firmly the forecaster believes in his/her expectations. In a dynamic environ-

ment like the stock market, the assertion that rational expectations can be improved via feedback is important because it opens avenues for further research comparing portfolio models utilizing adaptive expectations (historical data) with those utilizing rational expectations (subjective forecasts as inputs).

The second important finding of the study is that performance feedback is evidently superior to outcome feedback in improving the overall accuracy of forecasts. This finding may have critical implications for the training of forecasters in financial settings. Although not commonly used in depicting predictions of stock prices, probability distributions appear familiar to users of financial information. The use of probability distributions to forecast stock prices, if it is supported with training on subjective probability concepts, can potentially improve the investor's understanding and presentation of uncertainty in portfolio choice. This study suggests that such training may have a more pronounced impact if it is enhanced with feedback delineating various aspects of performance. The provision of such training may be viewed as an essential step towards establishing a new and effective channel of communication (via subjective probabilities) for disseminating financial knowledge and uncertainty. Further research about the use of probabilistic forecasting in different financial settings will be of great benefit to both the providers and the users of financial information.

### Acknowledgements

The authors gratefully acknowledge Wilpen Gorr, the associate editor, and the reviewers for their comments.

### Appendix: Performance measures

#### A.1. Probability score for multiple events

Let  $f = (f_1, \dots, f_m)$  be the forecast vector given by a forecaster for each of the stocks, with  $f_k$  denoting a probability forecast that the stock's

price change will fall into interval  $k$ ,  $k = 1, 2, \dots, m$ . Accordingly, let  $\mathbf{d} = (d_1, \dots, d_m)$  define an outcome index vector, with  $d_k$  taking on the value of 1 if the realized price change falls within interval  $k$ , and taking on the value of 0 if it does not fall within interval  $k$ . The probability score for multiple events (PSM) can then be defined as:

$$\text{PSM} = (\mathbf{f} - \mathbf{d})(\mathbf{f} - \mathbf{d})^T = \sum (f_k - d_k)^2.$$

Hence, the mean of probability scores ( $\overline{\text{PSM}}$ ) over a specified number of forecasting instances (i.e. over a given number of stocks) gives an index of a forecaster's overall accuracy level. The lower the score, the better the overall accuracy with respect to the stocks in question.

Components resulting from the Yates decomposition of the  $\overline{\text{PSM}}$  (Yates, 1988) are outlined next.

#### A.2. Calibration

Calibration provides information about the forecaster's ability to match the probability assessments with the realized relative frequencies. For example, suppose that for a set of 100 predictions, a forecaster assesses a probability of 0.4 that the given stock's price will increase more than 10%. This forecaster's 0.4 assessments are well-calibrated if an increase of more than 10% is actually observed on 40 of the 100 predictions. If the forecaster's other probability forecasts similarly match event frequencies, the forecaster is said to be well-calibrated.

Accordingly, a calibration score is a function of  $f_k$  (mean probability forecast for interval  $k$ ), and  $d_k$  (proportion correct for interval  $k$ ). In particular,

$$\text{Calibration} = \sum (f_k - d_k)^2.$$

Lower calibration scores indicate better performance in assigning appropriate probabilities to outcomes.

#### A.3. Mean slope

The mean slope index is another performance aspect that indicates the forecaster's ability to discriminate between instances when the actual

price change will and will not fall into the specified intervals. The higher the mean slope, the better the forecaster is able to discriminate. The mean slope is computed as follows:

$$\begin{aligned} \text{Mean slope} &= (1/m) \sum \text{Slope}_k \\ &= (1/m) \sum (\bar{f}_{1k} - \bar{f}_{0k}), \end{aligned}$$

where  $\bar{f}_{1k}$  is the mean of probability forecasts for a price change falling into interval  $k$ , computed over all the cases where the realized price change actually fell into interval  $k$ . Similarly,  $\bar{f}_{0k}$  is the mean of probability forecasts for a price change falling into interval  $k$ , computed over all the times when the realized price change did not fall into the specified interval.

#### A.4. Scatter

Scatter is that portion of the overall forecast variance that is not directly attributable to the forecaster's ability to discriminate between occasions when the actual price change will and will not fall into the specified intervals. Given that scatter basically reflects excessive variance, lower values are better. A scatter index is computed as follows:

$$\begin{aligned} \text{Scatter} &= \sum \text{Scatter}_k \\ &= \sum (1/N) [(N_{1k} * \text{Var}(f_{1k})) \\ &\quad + (N_{0k} * \text{Var}(f_{0k}))], \end{aligned}$$

where  $\text{Var}(f_{1k})$  is the conditional variance of the  $N_{1k}$  forecasts given for a price change falling into interval  $k$  when it actually occurred. Similarly,  $\text{Var}(f_{0k})$  is the conditional variance of the  $N_{0k}$  forecasts given for a price change falling into interval  $k$  when it did not occur. Obviously,  $N = N_{1k} + N_{0k}$ .

#### A.5. Forecast profile variance

Forecast profile variance summarizes the discrepancy between a forecaster's set of probabilities [i.e.  $\mathbf{f} = (f_1, \dots, f_m)$ ] and a uniform set of probabilities [i.e.  $\mathbf{f} = ((1/m), \dots, (1/m))$ ]. Accordingly, the forecast profile variance compares the

forecaster's probability profile with a flat profile that shows no variability across intervals. An index of the forecast profile variance could then be computed as

Forecast profile variance =

$$(1/N) \sum_n \left[ \sum_k (f_k - (1/m))^2 / m \right].$$

This measure provides an opportunity to examine the profiles of probability forecasts from an 'across-interval variance' point of view. It gives an index of how different the forecaster's probability set is from a set given by a uniform forecaster.

#### A.6. Skill

The combined effect of those  $\overline{\text{PSM}}$  components under the forecaster's control can be measured through a skill score, computed as

$$\begin{aligned} \text{Skill} &= \overline{\text{PSM}} - \sum \text{Var}(d_k) \\ &= \overline{\text{PSM}} - \sum [(\bar{d}_k) * (1 - \bar{d}_k)], \end{aligned}$$

where  $\text{Var}(d_k)$  gives the variance of the outcome index  $d_k$  for interval  $k$ . Given that the  $d_k$  values are determined by the price changes realized in the stock market,  $\sum \text{Var}(d_k)$  reflects an uncontrollable element of  $\overline{\text{PSM}}$  (the value of which is given by the conditions of the stock market). Subtracting this uncontrollable or 'base-rate' component from  $\overline{\text{PSM}}$ , we have, as a remainder, the global effect of all the accuracy components that are under the forecaster's control. Given that lower values of  $\overline{\text{PSM}}$  signify better accuracy levels, lower skill scores indicate better overall forecasting quality as exhibited by the probability forecasts.

#### References

- Adams, J.K. and P.A. Adams, 1961, Realism of confidence judgments, *Psychological Review* 68, 33–45.
- Adams, P.A. and J.K. Adams, 1958, Training in confidence judgments, *American Journal of Psychology* 71, 747–751.
- Ball, R. and P. Brown, 1968, An empirical evaluation of accounting income numbers, *Journal of Accounting Research* 6, 159–178.
- Beach, B.H., 1975, Expert judgment about uncertainty: Bayesian decision making in realistic settings, *Organizational Behavior and Human Performance* 14, 10–59.
- Benson, P.G. and D. Önkal, 1992, The effects of feedback and training on the performance of probability forecasters, *International Journal of Forecasting* 8, 559–573.
- Fama, E.F., 1965, The behavior of stock market prices, *Journal of Business* 38, 34–105.
- Fischer, G.W., 1982, Scoring-rule feedback and the overconfidence syndrome in subjective probability forecasting, *Organizational Behavior and Human Performance* 29, 352–369.
- Fischhoff, B. and D. MacGregor, 1982, Subjective confidence in forecasts, *Journal of Forecasting* 1, 155–172.
- Friedman, D., 1983, Effective scoring rules for probabilistic forecasts, *Management Science* 29, 447–454.
- Hogarth, R.M., 1975, Cognitive processes and the assessment of subjective probability distributions, *Journal of the American Statistical Association* 70, 271–289.
- Jensen, M.C., 1968, The performance of mutual funds in the period 1945–64, *The Journal of Finance* 2, 389–416.
- Kidd, J.B., 1973, Scoring rules for subjective assessments, Paper written for the Annual Conference of the Operational Research Society, Torbay, England.
- Lichtenstein, S. and B. Fischhoff, 1980, Training for calibration, *Organizational Behavior and Human Performance* 26, 149–171.
- Lichtenstein, S., B. Fischhoff and L.D. Phillips, 1982, Calibration of probabilities: The state of the art to 1980, in: D. Kahneman, P. Slovic and A. Tversky, eds., *Judgment Under Uncertainty: Heuristics and Biases* (Cambridge University Press, Cambridge).
- Moriarty, M.M., 1985, Design features of forecasting systems involving management judgments, *Journal of Marketing Research* 22, 353–364.
- Muradoglu, G. and T. Oktay, 1993, Calendar anomalies in the Turkish stock market, Presented at the International Conference on Business and Economic Development in Middle Eastern and Mediterranean Countries, Istanbul.
- Muradoglu, G. and D. Önkal, 1992, Semi-strong form efficiency in a thin market: A case study, Presented at the 19th European Finance Association Meeting, Lisbon.
- Muradoglu, G. and M. Unal, 1993, Weak-form efficiency in the thinly traded Istanbul Securities Exchange, Presented at the International Conference on Business and Economic Development in Middle Eastern and Mediterranean Countries, Istanbul.
- Murphy, A.H., 1972a, Scalar and vector partitions of the probability score: Part I. Two-state situation, *Journal of Applied Meteorology* 11, 273–282.
- Murphy, A.H., 1972b, Scalar and vector partitions of the probability score: Part II. N-state situation, *Journal of Applied Meteorology* 11, 1183–1192.
- Murphy, A.H., 1973, A new vector partition of the probability score, *Journal of Applied Meteorology* 12, 595–600.

- Murphy, A.H. and H. Daan, 1984, Impacts of feedback and experience on the quality of subjective probability forecasts: Comparison of results from the first and second years of the Zierikzee experiment, *Monthly Weather Review* 112, 413–423.
- Murphy, A.H., W. Hsu, R.L. Winkler and D.S. Wilks, 1985, The use of probabilities in subjective quantitative precipitation forecasts: Some experimental results, *Monthly Weather Review* 113, 2075–2089.
- Oskamp, S., 1962, The relationship of clinical experience and training methods to several criteria of clinical prediction, *Psychological Monographs* 76.
- Pickhardt, R.C. and J.B. Wallace, 1974, A study of the performance of subjective probability assessors, *Decision Sciences* 5, 347–363.
- Ronis, D.L. and J.F. Yates, 1987, Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method, *Organizational Behavior and Human Decision Processes* 40, 193–218.
- Schaefer, R.E. and K. Borcharding, 1973, The assessment of subjective probability distributions: A training experiment, *Acta Psychologica* 37, 117–129.
- Stael von Holstein, C.A.S., 1972, Probabilistic forecasting: An experiment related to the stock market, *Organizational Behavior and Human Performance* 8, 139–158.
- Winkler, R.L., 1969, Scoring rules and the evaluation of probability assessors, *Journal of the American Statistical Association* 64, 1073–1078.
- Wright, G. and P. Ayton, 1986, Subjective confidence in forecasts: A response to Fischhoff and MacGregor, *Journal of Forecasting* 5, 117–123.
- Yates, J.F., 1988, Analyzing the accuracy of probability judgments for multiple events: An extension of the covariance decomposition, *Organizational Behavior and Human Performance* 30, 132–156.
- Yates, J.F., L.S. McDaniel and E.S. Brown, 1991, Probabilistic forecasts of stock prices and earnings: The hazards of nascent expertise, *Organizational Behavior and Human Decision Processes* 40, 60–79.
- Yates, J.F., Y. Zhu, D.L. Ronis, D.-F. Wang, H. Shinotsuka and M. Toda, 1989, Probability judgment accuracy: China, Japan, and the United States, *Organizational Behavior and Human Decision Processes* 43, 145–171.

**Biographies:** Dilek ÖNKAL is an Assistant Professor of Decision Sciences at Bilkent University, Turkey. She received a Ph.D. in Decision Sciences from the University of Minnesota, and is doing research on decision analysis and probability forecasting. She has published in the *European Journal of Operational Research*, *International Forum for Information and Documentation*, *International Journal of Forecasting*, *Journal of Behavioral Decision Making*, and the *Journal of Forecasting*.

Gulnur MURADOĞLU is an Assistant Professor of Finance at Bilkent University, Turkey. She received a Ph.D. in Accounting and Finance from Bogazici University and is doing research on stock market efficiency and stock price forecasting. She has published in the *European Journal of Operational Research*, *International Journal of Forecasting*, *Journal of Forecasting*, and *The Middle East Business and Economic Review*.