ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES 49, 60–79 (1991)

Probabilistic Forecasts of Stock Prices and Earnings: The Hazards of Nascent Expertise

J. FRANK YATES

The University of Michigan

LINDA S. MCDANIEL

University of Washington

AND

ERIC S. BROWN

Columbia University

Undergraduate and graduate students in finance courses made probabilistic forecasts of the quarterly changes in the stock prices and earnings of publicly traded companies. Consistent with previous findings (Staël von Holstein, 1972), the overall accuracy of both price and earnings forecasts was very modest; subjects would have been more accurate had they predicted that price changes were equally likely to fall into any of the specified ranges. Also consistent with earlier suggestions of "inverted" expertise effects, undergraduate subjects were more accurate than graduate subjects. Decompositional analyses of subjects' judgments were consistent with the hypothesis that graduate students' relatively poor accuracy was affected by their greater tendency to report forecasts that varied from one stock to the next instead of the same forecast for every one. It is argued that the most plausible explanation is that the graduate subjects responded to cues they thought were predictive, but which actually were not. However, it cannot be ruled out completely that the graduate subjects attended to truly predictive cues, but were simply unable to use them appropriately.

In the early 1970s, Staël von Holstein (1972) performed an experiment concerning the stock market. He focused on the accuracy of stock price predictions. For each of 12 stocks, subjects made probabilistic forecasts that price changes over successive 2-week periods would fall into five

This research was supported in part by an award from the University of Michigan's College of Literature, Science, and the Arts. It is our great pleasure to acknowledge the incisive criticisms and helpful suggestions of Michael Doherty and Charles Gettys for an earlier version of this paper.

Requests for reprints should be sent to: J. Frank Yates, Department of Psychology, University of Michigan, 330 Packard Road, Ann Arbor, MI 48104-2994.

0749-5978/91 $3.00

specified intervals which partitioned the continuum. The intervals were defined by cut-points at −3%, −1%, +1%, and +3%. For instance, on a given occasion, the subject might report probability $p_1 = .30$ that Stock X would experience a fall in price of 3% or more, probability $p_2 = .25$ that the price change would be a decline of 1%–3%, probability $p_3 = .20$ of a change somewhere between a decline of 1% and an increase of 1%, and so forth, with the constraint that the probability judgments for all five intervals sum to 1.0. Staël von Holstein's primary aim was training. Every 2 weeks, he gave his subjects feedback about their accuracy to that point in the study. This feedback was generated by a quadratic scoring rule equivalent to the probability score for multiple events, as described by Yates (1988). Apparently, it was hoped that scoring rule feedback would improve forecasting performance.

The results of Staël von Holstein's experiment were surprising. First of all, the training was ineffective. Scoring rule feedback yielded virtually no improvement in accuracy over the 20 weeks the experiment lasted. Second, the subjects' predictions were remarkably *in*accurate. For example, only 3 of the 72 subjects' average scores were better than that of a uniform forecaster. A "uniform forecaster" is an individual who consistently behaves as if all the specified possibilities are equally likely. So a uniform forecaster would have reported a 20% chance that the actual price change for every stock would fall into each of the five ranges described by Staël von Holstein. Third, there was some evidence that the relationship between accuracy and expertise is almost the inverse of what many people would expect. The rank ordering of subject groups, in terms of accuracy, was: statisticians > stock market experts > university business students > university business *teachers* > bankers.

In the present research, we were not concerned with feedback training. Instead, we addressed three additional fundamental issues raised by Staël von Holstein's results. First, we set out to test the reliability of the previous findings on overall accuracy. Could probability judgments about securities *really* be as bad as the results suggested? As Staël von Holstein noted, it was plausible that certain aspects of his design might have degraded his subjects' performance, e.g., the 2-week forecasting horizon, which is apparently much shorter than customary for professional forecasters. Also, stock prices might be inherently an especially difficult quantity to predict. Perhaps judgments about other financial aspects of target companies, e.g., their earnings, would be better. Our second objective was to verify the suggested inverse expertise–accuracy relationship: Do novices indeed make financial forecasts better than more knowledgeable individuals? Finally, assuming that they did replicate, we sought explanations for the previous results: Why should security forecasts be so poor? And why should naive forecasters sometimes outperform sophisticated ones?

The conceptual framework which guided the study is illustrated in the panels of Fig. 1, which closely resemble lens model diagrams (Hammond, 1966). In our study, we examined the forecasting behavior of undergraduate and graduate finance students. We regarded the former as "novices" and the latter as "semi-experts." First consider Fig. 1(a), which depicts the situation we would anticipate for semi-experts. The actual price of a stock is indicated on the left. The individual's probabilistic forecast of that price is represented on the right as P′ (Price). The actual price is shown to be related to some collection of relevant cues, only a few of which will capture the forecaster's attention. On the other hand, semi-experts' judgments can be expected to be heavily affected by a host of irrelevant cues—signs the forecaster believes are predictive of price activity, but which really are not. This expectation is a generalization from other contexts. One such domain is medicine, where it is found that experienced physicians' diagnoses are consistently affected by signs and symptoms that have no reliable statistical relationships to patients' actual medical conditions (e.g., Poses, Cebul, Collins, & Fager, 1985). Similar effects have been observed among agricultural experts (Gaeth & Shanteau, 1984). Thus, as suggested at the top of Fig. 1(a), the relationship between actual prices and the semi-expert's forecasts should be very poor.

Now consider Fig. 1(b), which characterizes our view of how a naive subject can be expected to approach the task of predicting stock prices. As a novice, this individual knows little, and readily acknowledges this ignorance. Accordingly, he or she pays attention to few if any cues—relevant or irrelevant—when making forecasts. This implies judgments that are essentially constant, from one case to the next. For instance, for lack of any basis for doing otherwise, the novice might indicate that the probability of each stock increasing in price is his or her estimate of the percentage of all stock prices that increase during the first quarter of a year. Since such constant forecasts cannot covary with actual price changes, their accuracy is severely limited. Accuracy might suffer even further to the extent that the particular constant forecasts that are chosen are far from the pertinent base rates. On the other hand, constant forecasts enjoy the advantage of all conservative predictions. That is, as long as they are in the ballpark of the base rates, on average they are unlikely to be in error *dramatically*.

There is considerable evidence in the economic literature that stock prices are extremely difficult to anticipate. In fact, analyses have often indicated that stock price movements resemble a "random walk," with price changes in time period $t = T$ being essentially independent of changes in periods $t < T$ (e.g., Fama, 1965; Roberts, 1959). There have been demonstrations that this phenomenon is consistent with the hypothesis that the stock market is "efficient." That is, virtually instanta-

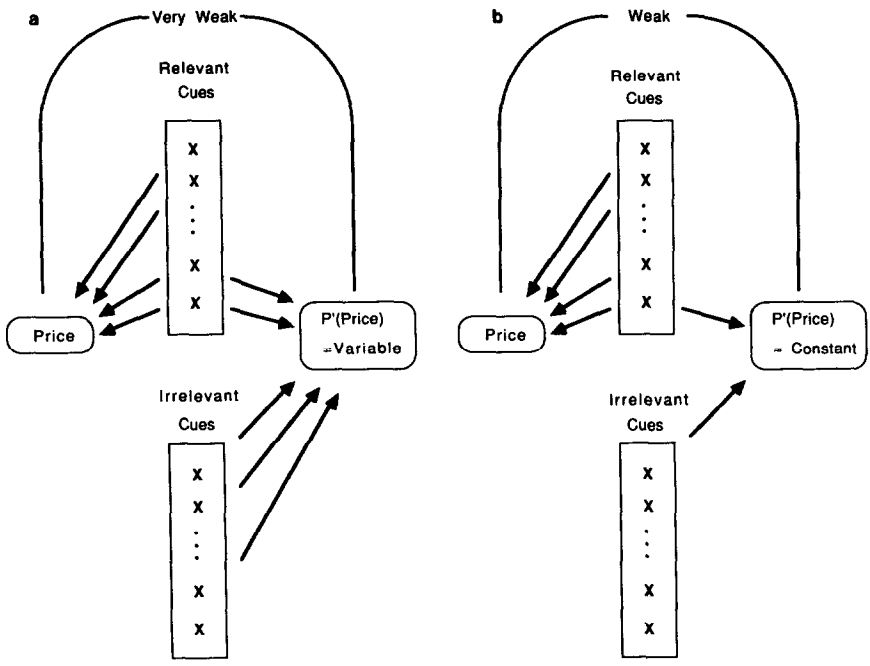

FIG. 1. Diagrammatic representation of the stock price forecasting situation: (a) For semi-experts, e.g., graduate students, (b) for novices, e.g., undergraduates.

neously, stock prices fully reflect all publicly available information about the companies offering those stocks (e.g., Samuelson, 1965). This is brought about by the aggregate actions of thousands of stock traders, whose purchases affect the demand for a stock and hence its price. Of course, this does not imply that *no one* can make good price forecasts. As an extreme counter-example, consider the case of individuals trading with inside information.

We hypothesized that forecasting firms' earnings per share would be easier than predicting their share prices. The price of a company's stock is largely a reflection of traders' expectations. These in turn are buffeted by a multitude of forces, many of which are unknowable, volatile, and outside the company's control, e.g., general market strength, rumors, and traders' beliefs about what other traders will do. In contrast, how much money a company makes is more directly determined by its own characteristics and actions (Lorie, Dodd, & Kimpton, 1985). Related to this argument is the fact that, when professional security analysts forecast a stock price, a major consideration that is taken into account is the firm's prospective earnings, model-based forecasts of which are published on a regular and frequent basis. But importantly, this is not the *sole* consider-

ation professionals use in predicting stock prices. Thus, the pervasive inaccuracy Staël von Holstein observed in probabilistic stock price forecasts should be less extensive for earnings forecasts.

METHOD

Subjects

All the subjects in the study were students. They were recruited through announcements in undergraduate and graduate finance classes at the University of Michigan Business School. Several inducements were offered to potential participants: (a) The opportunity to practice and evaluate their forecasting skills; (b) the chance to learn about probabilistic forecasting accuracy analysis; and (c) base and bonus monetary payment. Of the 31 individuals who eventually completed the study, 14 were undergraduate business administration majors and 17 were graduate students in business. Of the latter, all but one was studying for the MBA degree; the remaining participant was a Ph.D. student. The resulting two groups of "novices" and "semi-experts" were comparable in size to Staël von Holstein's (1972) subject groups.

Procedure

Each subject participated in an instruction and practice session. The subject was told that the task was to make probabilistic forecasts of the per-share stock prices and earnings of 31 companies listed on the New York Stock Exchange for the first quarter of 1986, which had begun 3 to 4 weeks previously. The companies had been randomly selected from those whose fiscal years end on December 31. The experimenter and the instruction booklet indicated that forecasts were to be made for percentage changes that might fall into six intervals which partitioned the continuum. The following hypothetical completed response form is the one included in the instructions:

INTERVAL	RANGE OF STOCK PRICE CHANGE	PROBABILITY
(6)	INcrease > +10%	15%
(5)	INcrease 5–10%	30%
(4)	INcrease up to 5%	20%
(3)	DEcrease 0 to 5%	30%
(2)	DEcrease 5–10%	5%
(1)	DEcrease > −10%	0%
	Total (Should be 100%):	100%

In the case of prices, each forecast was for the change from December 31 to March 31. Earnings are known to exhibit seasonal fluctuations. Thus, earnings forecasts were for changes from the first quarter in the previous year.

Every subject was provided with a folder which contained not only written instructions and response sheets, but background information about each of the target companies as well. This information included the name of the company, its industry, and its gross revenues for the most recent fiscal year available. The information sheet for a given company also listed its closing share price on the last day of each of the preceding eight quarters. In addition, it reported the earnings per share for each of those quarters, except the immediately preceding one, since these had not yet been published.

Subjects were told that each participant in the study would receive a token base payment of $5. The subject was also informed that an "accuracy score" would be computed for his or her forecasts. This score was in fact a linear transformation of the probability score for multiple events (Yates, 1988). It was emphasized that this score is "proper" (Winkler & Murphy, 1968). The printed instructions and the experimenter explained what this implies. That is, it was in the subject's best interests to report his or her true opinion about the probability of each price or earnings change, i.e., to avoid hedging. The subjects were promised a written report of the results of the study. The report would include an explanation of the scoring rule and the decompositional analysis applied to his or her judgments (see Yates, 1988, and the *Results and Discussion* section below). It would also include a ranked listing of all the participants' accuracy scores for both price and earnings forecasts, identified by codes. To gauge his or her own performance relative to the group, a given subject's report would indicate his or her code number. Finally, to provide a performance incentive and to reinforce the effect of the accuracy score's properness, subjects were told that the participant with the best combined accuracy score for price and earnings forecasts would receive a bonus of $30. The second, third, fourth, and fifth best forecasters were to receive bonuses of $20, $15, $10, and $5, respectively.

Subjects were allowed to take all the experimental materials home with them. To simulate naturalistic forecasting as well as possible, subjects were told that they were free to consult any information or individual they desired, other than another participant in the study. After they finished the forecasting task, subjects were to complete a postexperimental questionnaire. This instrument asked the subject to indicate his or her field of study, year in school, number of previous and current finance courses, and work experience related to finance. It also asked the subject to describe the sources consulted and the forecasting strategies used, to report the amount of time spent on the project, and to rate separately the diffi-

culty of the price and earnings forecasting tasks. Subjects were given a deadline of 10 days to complete their assignments.

RESULTS AND DISCUSSION

Subject Group Characteristics

As expected, the undergraduate subjects had taken fewer finance classes than their graduate counterparts, medians of three and five classes, respectively. Also, whereas only three of the 14 undergraduates (21.4%) reported at least some work experience related to finance, 10 of the 17 graduates did (58.8%). With respect to the forecasting exercises themselves, it is notable that there were no reliable group differences in the amount of time subjects reported spending and the rated difficulty of the tasks. Besides the information supplied by the experimenter, the major sources of company information the subjects consulted could be placed into three categories: (a) periodicals such as *The Wall Street Journal,* (b) Standard & Poor's stock reports, and (c) *The Value Line Investment Survey*. However, although only four of the 14 undergraduate subjects (28.6%) relied on two or more sources, 10 of the 17 graduate subjects did (58.8%).

Price Forecasts

Overall accuracy. The overall accuracy of subjects' forecasts was measured by the probability score for multiple events (Yates, 1988). This is a form of the quadratic scoring rule discussed by Brier (1950). It was also used by Staël von Holstein (1972) in his study of stock price forecasts. The scoring procedure can be described as follows.

For each stock, the subject reported a forecast vector $\mathbf{f} = (f_1, f_2, f_3, f_4, f_5, f_6)$, where f_k denotes a probabilistic forecast that the stock's price change will lie within interval I_k, $k = 1, \ldots, 6$, and where the intervals are those described in the procedure section. For each interval, there is an outcome index function d_k, which assumes the value 1 if the actual price change falls within that interval and the value 0 if it does not. Thus, we can also speak of an outcome index vector $\mathbf{d} = (d_1, d_2, d_3, d_4, d_5, d_6)$, defined in the obvious fashion. Intuitively, the outcome index can be seen as the forecast reported by a clairvoyant. A mortal forecaster's predictions are considered accurate to the extent that they approximate the outcome index. The "probability score for multiple events" is the scalar product of the difference between the forecast and outcome index vectors:

$$PSM(\mathbf{f},\mathbf{d}) = (\mathbf{f} - \mathbf{d})(\mathbf{f} - \mathbf{d})^T = \sum_{k=1}^{K} (f_k - d_k)^2, \qquad (1)$$

where , in the present case, $K = 6$. *PSM* ranges between 0 and 2. Clearly, the forecaster's aim should be to minimize its value. The mean value of the probability score over a given number of forecasting occasions ($\overline{PSM}$), provides a sense of the forecaster's characteristic accuracy level.

The first section of Table 1 indicates that the undergraduate subjects' price forecasts were significantly more accurate than those of their graduate counterparts. (All tests of accuracy statistics are nonparametric since the sampling distributions of these statistics are unknown.) However, none of the subjects was very accurate in absolute terms. This is most easily appreciated when their performance is compared to that of several hypothetical constant forecasters, who are often used as standards of comparison. Various measures for these forecasters are contained in Table 1 also.

The first standard is provided by a "uniform forecaster" who, as indicated previously, consistently reports that all of the specified events in a given situation are equally likely to occur. Thus, in the present situation, a uniform forecaster would assign probability $\frac{1}{6} \simeq .167$ to each of the six

TABLE 1

MEDIAN VALUES AND SIGNIFICANCE LEVELS (*p*) FOR VARIOUS ACCURACY MEASURES FOR UNDERGRADUATE (*UG*) AND GRADUATE (*GR*) SUBJECTS' PRICE (*P*) AND EARNINGS (*E*) FORECASTS, WITH CORRESPONDING MEASURES FOR UNIFORM (*U*), HISTORICAL (*H*), AND BASE RATE (*B*) FORECASTERS

Measure	Target	*UG*	*GR*	*p: UG* vs *GR*[a]	*U*	*H*	*B*
$\overline{PSM}$:	Prices	1.1538	1.1968	<.05	.8334	.7283	.5847
Overall acc.	Earnings	1.1000	1.2471	<.05	.8331	.7826	.7555
	p: P vs *E*[b]	<.05	—	—	—	—	—
Calib. index:	Prices	.3611	.3804	*ns*	.2487	.1436	.0000
Calibration	Earnings	.1149	.1285	*ns*	.0776	.0271	.0000
	p: P vs *E*	<.0001	<.0001	—	—	—	—
Mean slope:	Prices	−.0198	−.0163	*ns*	.0000	.0000	.0000
Covariation	Earnings	−.0141	−.0255	*ns*	.0000	.0000	.0000
	p: P vs *E*	*ns*	*ns*	—	—	—	—
Scat. index:	Prices	.1196	.2408	<.05	.0000	.0000	.0000
Poor cues/	Earnings	.1490	.2297	<.05	.0000	.0000	.0000
inconsist.	*p: P* vs *E*	*ns*	<.05	—	—	—	—
Profile var.:	Prices	.0381	.0627	<.05	.0000	.0024	.0414
Discrep. from	Earnings	.0335	.0653	<.05	.0000	.0289	.0129
uniform	*p: P* vs *E*	<.01	<.05	—	—	—	—
Skill index:	Prices	.5736	.6120	<.05	.2487	.1436	.0000
Overall,	Earnings	.3471	.4957	<.05	.0776	.0271	.0000
corrected	*p: P* vs *E*	<.0001	<.0001	—	—	—	—

[a] Mann-Whitney U test, one-tailed, *UG* predicted to be better.
[b] Median test, one-tailed, *E* predicted to be better.

intervals distinguished for subjects. The dotted line labelled "U" drawn through Fig. 2(a) illustrates graphically the distribution of those forecasts across all intervals. A second standard is provided by a "historical forecaster." This individual determines the relative frequency with which a given focal event has been noted to occur in the past. The historical forecaster then makes a forecast identical to that historical relative fre-

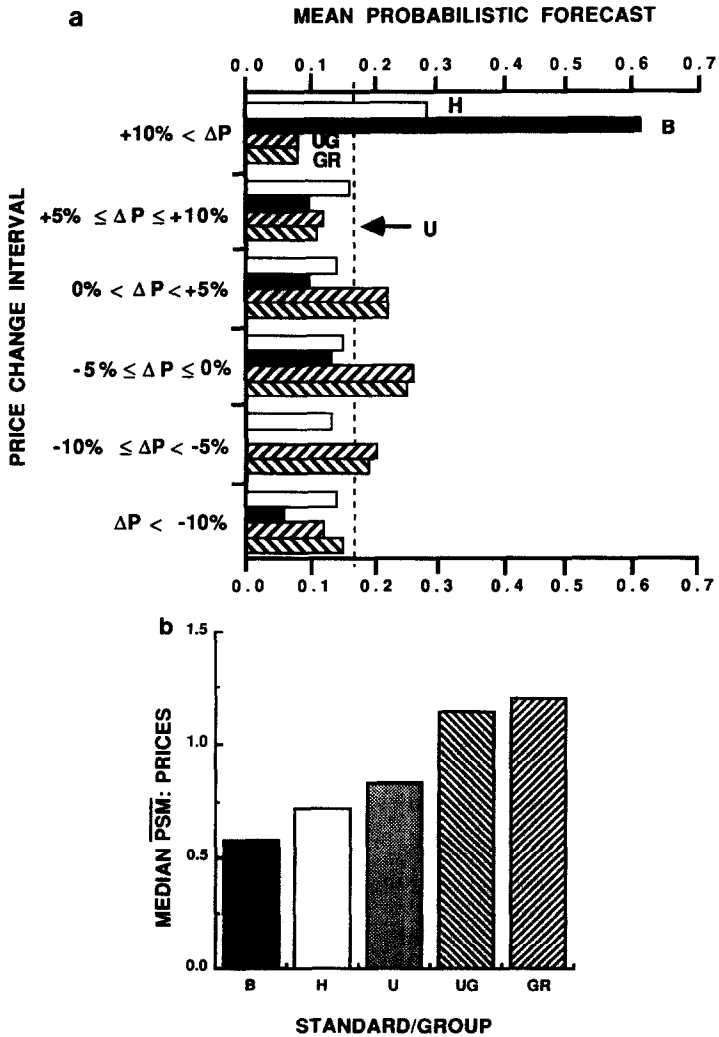

FIG. 2. (a) Mean price change forecasts for a uniform forecaster (*U*), a historical forecaster (*H*), a base rate forecaster (*B*), and the undergraduate (*UG*) and graduate (*GR*) subjects; (b) probability scores ($\overline{PSM}$) for uniform, historical, and base rate forecasters as standards and for the undergraduate and graduate subjects.

quency. For instance, if a subject chose to do so, he or she would have found that, in the background information provided about all the target companies, quarterly price gains more than 10% occurred about 28% of the time. Acting as a historical forecaster, for every company, that subject would have indicated a 28% chance of a price rise greater than 10%. The open bars in Fig. 2(a) describe the entire distribution of historical forecasts. A "base rate forecaster" provides a third important standard. This fictional forecaster can somehow anticipate the actual relative frequency, i.e., the base rate, with which a future event actually does occur. In the first quarter of 1986, about 61% of the 31 stocks subjects were asked to consider rose in price more than 10%. Accordingly, a base rate forecaster would have always reported a probabilistic forecast of .61 for a price increase over 10%. The black filled bars in Fig. 2(a) represent the complete distribution of base rate forecasts in this study.

It is easy to show that, regardless of what price changes might have occurred in the present study, a uniform forecaster was guaranteed to earn a mean probability score of $\overline{PSM} = .8334$.[1] As illustrated in Fig. 2(b), the accuracy of both the undergraduate and graduate subjects fell far short of that standard. Using decompositions of $\overline{PSM}$ into various accuracy components, it is straightforward to demonstrate that, using the previous price information provided in subjects' background folders, a historical forecaster would have earned a mean probability score of .7283. Similarly, we can show that a base rate forecaster would have done even better, earning a score of $\overline{PSM} = .5847$. As also shown in Fig. 2(b), both subject groups fell far short of the standards set by the historical and base rate forecasters.

An analysis note: Accuracy decomposition. Probabilistic forecasting accuracy is not a unitary construct. On the contrary, overall accuracy measures can be decomposed into several distinct elements (e.g., Murphy, 1972a, 1972b, 1973; Sanders, 1963). Such decompositions then point toward explanations of puzzling accuracy phenomena, e.g., the present subjects' pervasive inaccuracy and the greater forecasting success by the less experienced subjects. In general, accuracy decompositions can be conceptualized as follows:

$$\text{Accuracy} = f(\text{Uncontrollable Factors; Controllable Factors}) \quad (2)$$

That is, overall accuracy is a function of aspects of the forecasting task that are independent of the forecaster's actions as well as others which indeed are determined by what the forecaster does.

[1] Our subjects could not achieve precisely this value because they reported judgments to only the nearest whole percentage point.

In the case of the covariance decomposition of $\overline{PSM}$ as an overall accuracy measure, the conceptual partition is this (Yates, 1982, 1988):

$$\text{Accuracy} = f(\text{Uncontrollable: Base Rate; Controllable: Calibration, Covariation, Scatter}) \quad (3)$$

For instance, the base rate with which stock prices rise or fall is completely outside the forecaster's control. Nevertheless, it has an effect on his or her probability score. Thus, in interpreting how good or bad the forecaster's performance is, that effect should be removed. Or, equivalently, attention should be focused on those accuracy dimensions that *are* subject to the forecaster's influence—calibration, the covariation of prices and judgments, and forecast scatter. The meanings of these constructs are described in the next three subsections, which present the results of the decomposition analysis of subjects' price forecasts. More complete detail about the relevant statistics is provided by Yates (1988).

Calibration. Probabilistic forecasts are said to be well-calibrated to the extent that the average forecast matches the observed base rate. For example, suppose that, on each of 100 occasions, an analyst reports a forecast that the given stock will fall in price between 5% and 10%, and that the average of such forecasts is .20. Then, if the analyst's forecasts are perfectly calibrated, the prices for exactly 20 of the 100 stocks in fact will experience declines of 5%–10%. The patterned bars in Fig. 2(a) show the mean forecasts by the undergraduate and graduate subjects for each range of potential price changes. A comparison of those bars to the black base rate bars suggests that the subjects' calibration was not especially good. Perhaps more importantly, however, note that the bars for the undergraduate and graduate subjects differ only slightly from each other. This indicates that the calibration of the groups was comparable. Thus, differences in calibration cannot account for the groups' differences in overall accuracy.

The degree of calibration present in a collection of forecasts can be characterized numerically by the "calibration index," defined as follows:

$$CI = \sum_{k=1}^{K} (\bar{f}_k - \bar{d}_k)^2, \quad (4)$$

where the notational conventions are as before, and $\bar{f}_k$ and $\bar{d}_k$ are the mean forecast and base rate for Interval k, respectively. The calibration index section of Table 1 confirms the previous impression that, on average, the calibration of the undergraduate and graduate subjects' forecasts was essentially the same.

Covariation. Suppose a forecaster's opinions are good. Then high prob-

abilities should be assigned to a given price change interval on those occasions when the actual price changes in fact do fall within that interval, and low probabilities should be reported when the changes fall outside that interval. That is, the forecasts should be strongly related to the occurrences of eventual individual price changes. This aspect of judgment accuracy is embodied in the covariation between forecasts and outcome indexes. For a given Interval k, the statistical covariance of these quantities is given by

$$\text{Cov}(f_k, d_k) = \text{Slope}_k \text{Var}(d_k), \tag{5}$$

where

$$\text{Slope}_k = (\bar{f}_{1k} - \bar{f}_{0k}) \tag{6}$$

is the difference in the mean forecasts for a price change falling into Interval k when that actually does occur ($\bar{f}_{1k}$) and when it does not ($\bar{f}_{0k}$), and

$$\text{Var}(d_k) = \bar{d}_k(1 - \bar{d}_k) \tag{7}$$

is the variance of the outcome index d_k for that interval. Clearly, the slope is the covariance contributor that is under the forecaster's control and is thus of primary interest. Accordingly, the mean value of the slope across all intervals,

$$\text{Mean Slope} = (1/K) \sum_{k=1}^{K} \text{Slope}_k, \tag{8}$$

serves as a good summary of the forecaster's covariation skills.

The mean slope section of Table 1 indicates that the abilities of the undergraduate and graduate subjects to vary their forecasts with actual price changes did not differ from each other reliably. Thus, like calibration, covariation differences cannot explain the overall accuracy advantage exhibited by the undergraduates. It is also of some interest that the covariation skills of both groups were essentially nil, with the mean slope statistics being nonsignificantly different from zero, although they were slightly negative.

Scatter. In a sense, the converse of skill at covarying one's forecasts with actual occurrences is the ability to avoid varying one's predictions independently of those occurrences. Such independent variation is referred to as "scatter." It is most plausibly due to either or both of two factors. First, the forecaster might make predictions on the basis of weak cues, information that is thought to be reliably related to prices, for instance, but which actually is not. Alternatively, the forecaster might be

inconsistent in responding to whatever information he or she does take into account, e.g., sometimes making optimistic predictions on the basis of a cue and at other times pessimistic forecasts from the same cue.

The scatter that is present in forecasts of price changes in Interval k is represented numerically by the scatter index:

$$SI_k = (1/N)[N_{1k}\mathrm{Var}(f_{1k}) + N_{0k}\mathrm{Var}(f_{0k})], \tag{9}$$

where $\mathrm{Var}(f_{1k})$ is the conditional variance of the N_{1k} forecasts for a price change falling into Interval k when that actually occurred, and $\mathrm{Var}(f_{0k})$ is the corresponding conditional variance of such judgments for the N_{0k} occasions when it did *not* occur, $N = N_{1k} + N_{0k}$. That is, SI_k is analogous to error variance in the analysis of variance. The scatter index for the entire collection of forecasts, over all intervals, is simply the sum of those for the individual indexes:

$$SI = \sum_{k=1}^{K} SI_k \tag{10}$$

As indicated in the scatter section of Table 1, there was statistically reliably less scatter in the undergraduate subjects' forecasts than in those of the graduate subjects. Thus, this appears to be the immediate basis for the overall superior accuracy of the undergraduates' predictions.

Forecast profile variance. Why were the graduate subjects' forecasts plagued by more scatter than were the undergraduates' predictions? According to the hypothesis described in the introduction, graduate subjects should be expected to vary their forecasts in response to cues their previous instruction and practical experience suggested to be predictive of actual price changes. Undergraduate subjects should be less likely to do this, because they have had less instruction and experience and recognize the limits of their knowledge. The graduate subjects' strategy would be detrimental if they responded to irrelevant cues or if they did rely on relevant cues, but improperly. The observed difference in scatter is consistent with the proposed hypothesis. Further evidence is provided by examining the "profiles" of the subjects' forecasts across all six of the price change intervals that were specified.

Recall that, for any given stock, a uniform forecaster would make the following collection of forecasts for price changes in the respective intervals:

$$\mathbf{f} = (.167, .167, .167, .167, .167, .167)$$

That is, the profile of forecasts is "flat," with forecasts exhibiting no variability from interval to interval. The current hypothesis implies that

the undergraduates' forecast profiles should more closely resemble that of a uniform forecaster than should those of the graduate subjects. This in turn implies that the across-interval variances of those forecasts should be smaller for the undergraduates. The profile variance section of Table 1 shows that this in fact was the case. Staël von Holstein (1972, p. 151) noticed a similar tendency for his more accurate subjects to report more uniform forecasts.

Additional, indirect evidence compatible with the proposed hypothesis is contained in Table 2. Only one of the self-reported background variables had consistently high correlations with most of the price forecast accuracy measures: the number of finance classes taken by the subject. Observe in the correlation matrix shown in Table 2(a) that scatter, profile variance, and $\overline{PSM}$ all increased with the number of classes the subject had taken. Also note that profile variance was almost perfectly correlated with scatter and $\overline{PSM}$. So, graduate subjects not only eschewed uniform forecasts, varying their judgments from one interval to the next. They also varied those nonuniform forecasts from one stock to another, but in a manner largely unrelated to actual stock prices. Table 2 shows that the number of classes was *inversely* related to the mean slope. It is as if classroom experience inspired the subjects to try their hands at various prediction strategies, but that either the strategies were "backwards" or the subjects' application of them was.

Earnings

The previously noted characteristics of subjects' price forecasts were completely paralleled in their earnings forecasts. Thus, as shown in Fig. 3(a), the mean earnings forecast patterns of the undergraduate subjects were very similar to those of the graduate subjects, but the patterns of both subject groups were markedly different from the patterns of histor-

TABLE 2
CORRELATION MATRIX FOR NUMBER OF FINANCE CLASSES TAKEN AND VARIOUS PRICE FORECAST ACCURACY MEASURES

Measure	Classes	$\overline{PSM}$	Calib. index	Mean slope	Scat. index	Profile var.
Classes	—					
$\overline{PSM}$	.47**	—				
Calib. index	−.08	.28	—			
Mean slope	−.49**	−.43*	.44*	—		
Scat. index	.49**	.91**	−.06	−.46**	—	
Profile var.	.42**	.96**	.18	−.38*	.96**	—

* $p < .05$, two-tailed test.
** $p < .01$, two-tailed test.

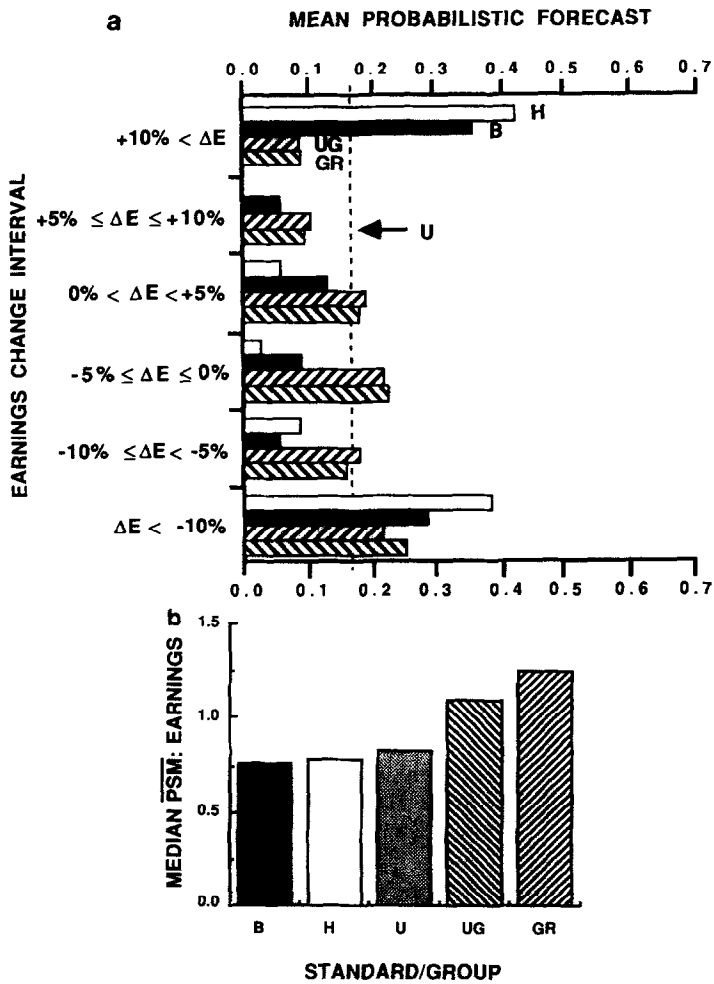

FIG. 3. (a) Mean earnings change forecasts for a uniform forecaster (*U*), a historical forecaster (*H*), a base rate forecaster (*B*), and the undergraduate (*UG*) and graduate (*GR*) subjects; (b) probability scores ($\overline{PSM}$) for uniform, historical, and base rate forecasters as standards and for the undergraduate and graduate subjects.

ical and base rate earnings forecasters. Also, as indicated in Fig. 3(b), the overall accuracy of subjects' earnings forecasts was significantly inferior to that of uniform, historical, and base rate forecasters. Further, as shown in Table 1, the undergraduates' earnings forecasts were more accurate than those of the graduate students. This difference apparently was due to the tendency of graduate subjects to vary their forecasts in such a way that they resulted in greater scatter.

Prices vs Earnings

Table 1 includes comparisons of various accuracy measures for price and earnings forecasts, along with the results discussed previously. In the $\overline{PSM}$ section we see that, as anticipated, the overall accuracy of the earnings forecasts was superior to that of the price forecasts, but only for the undergraduate subjects. Overall accuracy comparisons of price and earnings forecasts would be "unfair" if the base rates for prices and earnings were different. This is because, as indicated previously, $\overline{PSM}$ is affected by the base rates despite the fact that the forecaster has no control over those rates. As the solid bars in Figs. 2(a) and 3(a) indicate, the base rates for prices and earnings actually did differ.

Decompositions of $\overline{PSM}$ imply that it can be "corrected" by removing the contributions of the base rates (Yates, 1988). This results in what can be called a forecasting "skill index,"

$$\text{Skill Index} = \overline{PSM} - \sum_{k=1}^{K} \text{Var}(d_k), \tag{11}$$

which is, in fact, affected mainly by factors under the forecaster's control. As a virtual statistical necessity, given the previous analyses, Table 1 indicates that the skill indexes of undergraduate subjects' price and earnings forecasts were superior, i.e., lower, than the corresponding indexes for their graduate counterparts. Of more immediate significance, however, it is shown that both undergraduates and graduates exhibited greater skill at forecasting earnings rather than prices. Precisely how did the accuracy of earnings forecasts surpass that of price forecasts? Table 1 indicates that the major determinant of this superiority was the better calibration of the earnings forecasts. In view of the determinants of earnings as compared to prices, it is surprising that the observed differences in slope and scatter were weak and inconsistent.

Do individuals display a significant degree of consistency in their forecasting tendencies in the separate domains of prices and earnings (cf. Ronis & Yates, 1987)? The correlations in Table 3 suggest that, in most respects, they do. Subjects who made superior earnings forecasts tended to make better price predictions, too. Similar conclusions apply to calibration and scatter tendencies also, but not slope.

GENERAL DISCUSSION

Consistent with Staël von Holstein's (1972) findings, and under conditions somewhat closer to naturalistic ones, subjects' probabilistic forecasts of stock prices were shown to be surprisingly inaccurate. As expected, predictions of company earnings were better, though still not very

TABLE 3
PRICE VS EARNINGS CORRELATIONS (r), WITH SIGNIFICANCE LEVELS (p), FOR VARIOUS ACCURACY MEASURES

	Undergraduates		Graduates	
Measure	r	p	r	p
PSM	.74	<.005	.93	<.001
Calibration index	.75	<.005	.53	<.05
Mean slope	−.02	*ns*	.11	*ns*
Scatter index	.77	<.005	.85	<.001
Skill index	.74	<.005	.93	<.001
Profile variance	.74	<.005	.92	<.001
Difficulty rating	.25	*ns*	.09	*ns*

accurate in an absolute sense. This difference was achieved via better calibration for the earnings forecasts. Also in agreement with conclusions suggested by Staël von Holstein's group comparisons, undergraduates' predictions were more accurate—actually, less *in*accurate—than those of graduate students. Detailed analyses indicated that the undergraduates' relative advantage was due to their forecasts containing less variability that was independent of actual price and earnings activity. In turn, this was at least partly explained by the closer resemblance of their judgments to those of a uniform forecaster.

There are, of course, numerous caveats that preclude any sweeping conclusions that might be drawn from the present study. One is that the subjects were only students rather than practicing finance professionals (although most of the graduate subjects in fact did possess some professional experience). Moreover, forecasts were made for only a single financial quarter. Nevertheless, the results at least suggest several hypotheses that should be pursued in future work.

Some previous research on expertise has emphasized that experts' representations of problem situations tend to be more abstract than novices' representations, being built around underlying principles rather than surface features of the stimuli (e.g., Chi, Feltovich, & Glaser, 1981). Several of these studies have shown how such representational differences can sometimes lead to novices outperforming experts (e.g., Adelson, 1984; Chase & Simon, 1973). These anomalies occur because the experts are induced to try to apply their normally more effective representational structures to situations in which they are actually inappropriate, e.g., when a chess master is asked to memorize a chess board configuration that he is told comes from a real chess game but is actually randomly generated.

The superior performance of the less experienced subjects in the

present study does not appear to have its basis in the abstract vs concrete distinction. Instead, it seems grounded in another phenomenon that is sometimes noted in expertise research, that experts' representations are richer than those of novices (e.g., Murphy & Wright, 1984). In particular, recall that the forecasts of our presumed semi-experts, i.e., graduate student subjects, were more responsive to various differences among target companies than were the predictions of our novices, i.e., undergraduate subjects. Besides simply providing another illustration, the present results also highlight two hazards of the enrichment phenomenon. These hazards implicitly have been acknowledged in the judgment literature, but not in most previous expertise research.

As a person acquires more experience within a domain, he or she also forms more beliefs about which cues are predictive of the relevant target events. In "simple" natural systems, false beliefs are relatively easily corrected via feedback, as illustrated by the fact that most children are rapidly disabused of many of their misconceptions about mechanics (see, for example, Kaiser & Proffitt, 1984). But in more complex systems, such as those encountered in medicine and business, focal events are so overdetermined by multiple causes that it is virtually impossible for a person to rely on unaided observation to correct his or her erroneous beliefs. Accordingly, weak cues can continue indefinitely to be added to those that affect judgments, e.g., stock price forecasts. The result is that, although we might expect that greater experience will lead to demonstrably greater accuracy, it instead simply results in more useless variation in judgments, e.g., scatter (cf. Gaeth & Shanteau, 1984; Poses et al., 1985).

Even if the cues that are gradually added to the pool considered by a judge are valid, this does not guarantee that they will enhance the judge's performance. Lens model research (e.g., Dudycha & Naylor, 1966) indicates that the addition of such cues can be detrimental in two ways. First, the accuracy of judgments can decline because the judge misuses the additional cues. Second, the new cues make the judgment task more difficult, and hence diminish the judge's reliability.

One of our referees was skeptical about the prediction that semi-experts' forecasts would exhibit greater scatter because those individuals would take more cues into account than would novices. It was suggested that a reasonable alternative prediction is that attending to more cues would produce *less* scatter. The argument is as follows: Suppose any judge effectively averages the values of all the cues that he or she uses. In terms of Figs. 1(a) and 1(b), it is as if, intervening between the individual cues and the process by which the person generates a judgment, those individual cues are reduced to a composite, average cue. Then, because of the phenomena described in such principles as the central limit theorem, the composite cues based on large numbers of cues should be less

variable than those that synthesize fewer ones. The present results argue against such a process, or at least such a process being executed very well. Instead, it appears that, for judges at the expertise level of our subjects, trying to cope with additional cues is overwhelming.

In everyday practice within the financial community, probabilistic forecasts are virtually unknown. Instead, predictions typically are reported as point estimates or unqualified forecasts of ranges. For instance, an analyst will simply assert that "Company X will earn $1.03 per share during the third quarter" or that "Company Y's earnings for this year should be between $3.50 and $3.80." Thus, it is difficult to make direct comparisons between the present results and those found in most previous studies of financial forecasting. However, parallels do exist, and some of the present conclusions can be expected to generalize. Moreover, a good case can be made that, although probabilistic forecasts are not routinely used in the financial world, they should be. At a minimum, they would allow the consumer of financial forecasts to know, for example, *how* firmly an analyst holds his or her expectation that Company Y's earnings will be in the $3.50–$3.80 range. Going even further, an investor could incorporate the analyst's probability judgments into financial decision making algorithms similar to those used in decision analysis.

In the present study, as in Staël von Holstein's (1972), subjects' probabilistic forecasts were worse than those of a uniform forecaster. This necessarily implies that they were also worse than those of a historical forecaster. There is evidence that professional forecasters' point estimates of earnings sometimes are more accurate than estimates generated by statistical models that rely on historical data. However, such advantages are far from universal (Armstrong, 1983). Thus, the present results do not seem at odds with "mainstream" findings on the accuracy of deterministic, i.e., nonprobabilistic, financial forecasts. This conclusion is more directly supported by an examination of professional forecasters' predictions for the cases used in this study. Every month, the "Institutional Brokerage Estimate System" publishes averages of leading professional analysts' forecasts of various statistics about securities. The January 16, 1986, report (which was, incidentally, readily available to our subjects) included the first quarter 1986 earnings forecasts for the 31 companies included in our study. The correlation of .22 between the consensus forecasts and the eventual actual earnings was positive but modest.

REFERENCES

Adelson, B. (1984). When novices surpass experts: The difficulty of a task may increase with expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **10**, 483–495.

Armstrong, J. S. (1983). Relative accuracy of judgmental and extrapolative methods in forecasting annual earnings. *Journal of Forecasting*, **2**, 437–447.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**(1), 1–3.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, **4**, 55–81.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, **5**, 121–152.

Dudycha, L. W., & Naylor, J. C. (1966). Characteristics of the human inference process in complex choice behavior situations. *Organizational Behavior and Human Performance*, **1**, 110–128.

Fama, E. F. (1965). The behavior of stock market prices. *Journal of Business*, **38**, 34–105.

Gaeth, G. J., & Shanteau, J. (1984). Reducing the influence of irrelevant information on experienced decision makers. *Organizational Behavior and Human Performance*, **33**, 263–282.

Hammond, K. R. (1966). Probabilistic functionalism: Egon Brunswik's integration of the history, theory and method of psychology. In K. R. Hammond (Ed.), *The psychology of Egon Brunswik*. New York: Holt, Rinehart, & Winston.

Kaiser, M. K., & Proffitt, D. R. (1984). The development of sensitivity to causally relevant dynamic information. *Child Development*, **55**, 1614–1624.

Lorie, J. H., Dodd, P., & Kimpton, M. H. (1985). *The stock market: Theories and evidence* (2nd ed.). Homewood, IL: Dow Jones-Irwin.

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, G. L., & Wright, J. C. (1984). Changes in conceptual structure with expertise: Differences between real-world experts and novices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **10**, 144–155.

Poses, R. M., Cebul, R. D., Collins, M., & Fager, S. S. (1985). The accuracy of experienced physicians' probability estimates for patients with sore throats. *Journal of the American Medical Association*, **254**, 925–929.

Roberts, H. V. (1959). Stock market 'patterns' and financial analysis: Methodological suggestions. *Journal of Finance*, **14**, 1–10.

Ronis, D. L., & Yates, J. F. (1987). Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior and Human Decision Processes*, **40**, 193–218.

Samuelson, P. A. (1965). Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review*, **6**, 41–49.

Sanders, F. (1963). On subjective probability forecasting. *Journal of Applied Meteorology*, **2**, 191–201.

Staël 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., & Murphy, A. H. (1968). "Good" probability assessors. *Journal of Applied Meteorology*, **7**, 751–758.

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

Yates, J. F. (1988). Analyzing the accuracy of probability judgments for multiple events: An extension of the covariance decomposition. *Organizational Behavior and Human Decision Processes*, **41**, 281–299.

RECEIVED: September 1, 1988