ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Organizational Behavior and Human Decision Processes 91 (2003) 169–185

ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES

www.elsevier.com/locate/obhdp

Professional vs. amateur judgment accuracy: The case of foreign exchange rates[☆]

Dilek Önkal,[a,*] J. Frank Yates,[b] Can Simga-Mugan,[a] and Şule Öztin[c]

[a] *Faculty of Business Administration, Bilkent University, Ankara 06533, Turkey*
[b] *Business School or Psychology Department, University of Michigan, Ann Arbor, MI 48109, USA*
[c] *Sümerbank A. Ş., Atatürk Bulvarı, No: 70, Ankara 06533, Turkey*

Abstract

Highly knowledgeable people often fail to achieve highly accurate judgments, a phenomenon sometimes called the "process-performance paradox." The present research tested for this paradox in foreign exchange (FX) rate forecasting. Forty professional and 57 sophisticated amateur forecasters made one-day and one-week-ahead FX predictions in deterministic and probabilistic formats. Among the conclusions indicated by the results are: (a) professional accuracy usually surpasses amateur accuracy, although many amateurs outperform many professionals; (b) professionals appear to achieve high proficiency via heavy reliance on predictive information (unlike what has been observed before, e.g., for stock prices); (c) forecast format strongly affects judgment accuracy and processes; and (d) apparent overconfidence can transform itself into underconfidence depending on when and how forecasters must articulate their confidence.

© 2003 Elsevier Science (USA). All rights reserved.

Almost every practical decision, by individuals or organizations, rests at least partly on judgments concerning particular facts or occurrences. Picture a medical patient and her physician deciding whether to treat her illness by surgery rather than drugs. That choice undoubtedly is driven to some degree by their beliefs about the relative chances that the alternative treatments will relieve her condition, that those treatments will cause various side effects, and that she will react badly to those effects. Or imagine a company's board of directors deliberating a radical shift in the company's strategic direction. Quite plausibly, a key reason for even considering such a move is management's belief that markets for the company's core products are about to change markedly. How well decisions turn out depends on a host of considerations. But central to the eventual adequacy of those decisions is the accuracy of the judgments on which they are predicated. That is, judgment accuracy imposes a ceiling on decision quality. Thus, if their judgments about surgery outcomes tend to be highly inaccurate, then patients and physicians will often find themselves electing surgeries that leave the patients worse off than they would have been with drug treatments, and vice versa. Similarly, corporate boards whose strategic decisions repeatedly are grounded in erroneous predictions about future market trends cannot help driving their companies into insolvency. It is therefore essential that deciders—individuals and organizations—do whatever they can to assure that the judgments informing their decisions are as accurate as is possible and feasible.

When confronted with a significant decision problem, deciders could rely solely on their own judgments. Alternatively, they could consult with others, including professionals who, in effect, sell their assessments for a price. Thus, a patient and her physician might solicit the prognoses of a recognized authority on a treatment under consideration. Or a board of directors might commission a respected expert to draw on her knowledge and skills to render an informed opinion about how the markets for the company's products are likely to develop in the years ahead. This prospect of professional consultation brings to the fore several important practical and scientific questions.

[☆] Helpful comments on earlier versions of this article by Terry Connolly and two anonymous reviewers are gratefully acknowledged.

[*] Corresponding author.

E-mail addresses: onkal@bilkent.edu.tr (D. Önkal), jfyates@umich.edu (J.F. Yates), simga@bilkent.edu.tr (C. Simga-Mugan).

0749-5978/03/$ - see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S0749-5978(03)00058-X

The first question is basic indeed: Just how accurate should deciders anticipate that professionals' judgments would be? Should they expect the accuracy of those judgments to be so much better than the accuracy of their own "amateur" judgments as to justify their expense? Or would the deciders be better off simply relying on their personal assessments? It is hard to imagine that, in virtually any domain, professional judgment would not generally be significantly superior to that of amateurs. After all, it seems reasonable to expect that professionals with deficient skills would be driven from the marketplace. Contrary to this sensible expectation, however, there have been numerous demonstrations of what is sometimes called a "process-performance paradox" (Camerer & Johnson, 1997), whereby individuals with vast knowledge about a domain nevertheless are unable to render highly accurate predictions in that arena (e.g., Enis, 1995; Spence & Brucks, 1997). So the answer to the "expectations question" is not a given.

Now, suppose that deciders do choose to solicit professional judgments to help guide their decisions. There are numerous formats in which they might request that the professionals deliver their opinions. An especially important format distinction is that between categorical or deterministic judgments, on the one hand, and probabilistic judgments, on the other. For instance, a consulting specialist might be asked for his "best guess" as to whether, in a case under consideration, a surgical procedure would succeed or fail. Alternatively, the consultant might be asked to indicate what is, in his view, the 0–100% probability that the operation would be a success. There are conceptual as well as practical reasons to favor probabilistic over categorical formats. Perhaps the most compelling is that judgments in probabilistic form allow deciders to trade off the actual uncertainty that always exists—whether acknowledged or not—against the significance or value of potential outcomes, as in expected utility operators (cf., Yates, Price, Lee, & Ramirez, 1996). Nevertheless, other considerations, including people's greater familiarity with them, might argue for deterministic judgments as the answer to the "format question" in a given decision situation.

Let us say that, for a particular domain, there are reliable differences in the accuracy of judgments offered by professionals and amateurs. It is then essential to achieve both contemporaneous and developmental explanations for those differences. A "contemporaneous explanation" would identify the alternative routes by which the professionals and amateurs arrive at their assessments in the here and now. That is, it would isolate and document specific differences in professionals' and amateurs' judgment processes that contribute to their accuracy differences. A "developmental explanation" would go a step further. It would shed light on why those process differences came to exist, e.g., particular training and work experiences. Implicit in both levels of explanation would be prescriptions for how managers could more readily identify or accelerate the development of true judgment expertise.

The research described in the present article sought answers to the above expectations, format, and explanation questions about professional judgment in a domain that has considerable significance in its own right, the domain of foreign exchange (FX) rates. Virtually every sizable organization today, from commercial enterprises to non-profit professional associations, must contend with rapid and relentless globalization, whether they like it or not. International currency differences are a key element of the globalization challenge. A shift in FX rates can mean that essential company supplies that were easily afforded last week are suddenly crushingly expensive. Or it could mean the opposite, that whereas the company's products were priced out of certain markets yesterday, they are competitive in those locales today. The challenge is felt at the personal level, too. Every consumer (or worker) is aware from newscasts that, because of FX rate shifts ("The dollar fell sharply today. . ."), the prices of the goods she must buy (or her company might sell abroad) can change suddenly, with immediate and often dramatic impact. Her paycheck effectively grows or shrinks in a flash (to nothing if her company fails). Obviously, organizations and individuals should seek to make decisions that protect and promote their interests in the face of potential FX rate changes. And clearly, the ability to accurately anticipate those changes allows for superior choices.

The importance of FX rates, as well as accurately predicting them, is even greater in some countries than others. Such rates assume special significance in economies undergoing rapid and volatile change. For in those circumstances, the value of the local currency relative to foreign currencies is likely to be highly changeable, too. This means that FX-related decisions made by organizations and consumers in such environments are particularly critical. And, therefore, so is the importance of accurate FX rate predictions. The empirical work reported in this article was conducted in Turkey, during a time when the Turkish economy had the characteristics described here. Circumstances were even more interesting because Turkish law permitted the free use of foreign currencies in everyday consumer transactions "on the street." Thus, it would not be unusual for an ordinary Turkish citizen to collect his pay in the local currency and then go to a currency exchange to convert it to the foreign currency he expected to fare best in the FX market in the future.

The conditions in Turkey thus provided an unusually rich opportunity to address the expectations, format, and explanation questions sketched above, where they really mattered. Specifically, these were the questions pursued:

Expectations. Should amateurs expect local professionals to make FX rate predictions that are significantly more accurate than their own, such that it would make sense for them to contract for the services of those professionals? Put another way, is it reasonable to expect professional FX forecasters in situations like that in Turkey to display the high degree of accuracy characteristic of, say, weather forecasters (e.g., Murphy & Brown, 1984)? That accuracy is often attributed to the fact that, besides having access to good models, weather forecasters make many, many judgments and they get immediate feedback about every one of them. Professional FX forecasters in Turkey make lots of judgments, too, although not as many as weather forecasters. They also have immediate feedback, even though it is doubtful that most of them analyze that feedback as carefully and systematically as weather forecasters do. On the other hand, there is little evidence that even experienced professionals can make highly accurate predictions of stock prices and earnings (cf., Staël von Holstein, 1972; Yates, McDaniel, & Brown, 1991). One proposed explanation for this modest level of accuracy, for stock prices, at least, is that the markets for such securities are efficient. This implies that, unless he is an insider in every company—which is, of course, impossible—even the most proficient professional cannot acquire facts that are inherently capable of supporting great accuracy over an extended time period. Similar arguments in FX markets have motivated claims that FX rates follow random walks, implying that it would be quite difficult for anybody to make rate predictions at above-chance accuracy levels (e.g., Mussa, 1979). But findings contradicting the random-walk view have also been reported (e.g., Lai & Pauly, 1992), along with various models predicting forecasting performance consistent with those reports (e.g., Sarantis & Stewart, 1995). Thus, a priori, it is by no means obvious what accuracy expectations ought to be.

Format. If one were to acquire FX rate predictions from professionals in the Turkish type of situation, would it matter how those judgments were rendered? As noted before, in principle, probabilistic judgments are preferable if for no other reason than that they would allow deciders to trade off uncertainty against other considerations. But when those professionals express their uncertainty probabilistically, is this truly informative? Or, in effect, do demands for probabilistic expression (and maybe other modes) simply add useless fuzziness to their deterministic judgments, perhaps because forecasters are unaccustomed to or fundamentally incapable of skillfully articulating their opinions that way?

Explanation. Suppose that professional and amateur FX rate judgments in the Turkish kind of context differ in accuracy. Why might that be so? At the level of contemporaneous explanations, the differences might find their origins in the judgment process variations implicated by decompositions of judgment accuracy that have received attention in recent years (e.g., Yaniv & Foster, 1995; Yates, 1994, 1998). By their nature, as snapshots of current processes, such analyses cannot definitively establish how process variations developed. But they can narrow the possibilities significantly.

Method

Participants

Forty FX dealers and business professionals responsible for FX forecasts for their companies served as the professionals in the study. The amateurs were 57 business students at Bilkent University in Ankara. Thus, the amateurs were not at all naïve. They generally had formal training in finance and forecasting and were all accustomed to the kind of personal, "street-level" currency trading common in Turkey at the time. The professionals were recruited through personal contact, the students via announcements in classes. All participants volunteered their services; they received no financial compensation.

Forecasting tasks

Each participant made forecasts for 10 exchange rates. There were five major "cross rates": US dollar/Deutschemark, British pound/US dollar, US dollar/Swiss franc, US dollar/Japanese yen, and Deutschemark/Japanese yen. And there were five "domestic-currency-based rates": US dollar/Turkish lira, Deutschemark/Turkish lira, British pound/Turkish lira, Swiss franc/Turkish lira, and Japanese yen/Turkish lira. Every participant made six sets of 50 forecasts for these rates, implying six basic tasks distinguished by formats and horizons:

- Point forecasts (Tasks 1 & 2): one-day horizon (daily), one-week horizon (weekly).
- Directional forecasts (Tasks 3 & 4): one-day horizon (daily), one-week horizon (weekly).
- Interval forecasts (Tasks 5 & 6): one-day horizon (daily), one-week horizon (weekly).

The term "horizon" refers to the time between when the forecaster made a prediction and the future date to which that prediction referred:

One-day horizon. For a daily or one-day-ahead prediction, the participant's aim was to anticipate the Reuters 11 a.m. TLFX-FY rate for the following day, the standard in the FX financial community. As background information, for a given currency, every participant was provided with the daily rate values for the previous four months (78 trading days in total) in graphical form, and also the most recent 10 values, in tabular form. For each of the 10 rates, each participant made one-day-ahead forecasts for five consecutive days, for a total of 50 daily forecasts.

One-week horizon. For a weekly or one-week-ahead forecast, the participant sought to predict the Reuters 11 a.m. TLFX-FY rate for Monday one week hence, the opening rate for the week. As background, for each currency, the participant was provided with the weekly Monday-opening values for the previous 18 months (78 Monday openings in total), again in graphical form, as well as the last 10 values in a table. For each of the 10 rates, each participant made one-week-ahead forecasts for five consecutive weeks, for a total of 50 weekly forecasts.

Essential requirements of the alternative prediction formats were as follows:

Point forecasts: Provide a single-value prediction for the rate in question.

Directional forecasts: First indicate whether the focal rate will either (a) increase or (b) decrease or remain the same. Then state a 50–100% probability judgment that the indicated directional prediction will indeed prove to be correct.

Interval forecasts: Specify an interval for the focal rate such that there is a 90% probability that the true value of that rate will in fact be captured by that interval, i.e., a 90% credible interval (cf., Yates, 1990, pp. 21–23).

Appendix A presents the specific instructions given to participants. It also shows an illustration of the form participants used to render their judgments.

Personal performance expectations

After receiving instructions but before reporting judgments, each participant was asked to state personal performance expectations as follows:

Point forecasts: "You will be making 50 (daily/weekly) point forecasts in total. In how many cases (out of 50) do you expect the realized value to be exactly equal to your point forecast?"

Directional forecasts: "You will be making 50 (daily/weekly) directional forecasts in total. In how many cases (out of 50) do you expect the realized change to fall in the direction you predicted?"

Interval forecasts: "You will be making 50 (daily/weekly) interval forecasts in total. In how many cases (out of 50) do you expect the realized value to fall within your prediction interval?"

Procedure

At the beginning of the first session, participants were given detailed information about the study. Forecast elicitation formats were explained and examples were given. Participants were informed that various performance scores would be computed for their individual forecasts. They were also told that no information about their predictive accuracy would be disclosed to other participants (or, in the case of the professionals, to their managers or co-workers). After that, participants reported their initial sets of forecasts. They did the same for subsequent daily and weekly sessions.

Results and discussion

The findings are organized according to the three kinds of forecasting formats used by the participants: point, directional, and interval. In each instance, we discuss the implications of the results for the basic questions set out initially, the expectations, format, and explanation questions.

Point forecasts

One of the most commonly used measures of overall point judgment accuracy is the median absolute percentage error (*MdAPE*; cf., Armstrong & Collopy, 1992). Its 0–100 range is one of its primary attractions, for it permits easy comparisons across quantities that have radically different scales, as in the present study, e.g., US dollars vs. Japanese yen vs. Turkish lira. Thus, *MdAPE* was the index employed here to evaluate the accuracy of our participants' point forecasts of FX rates. The absolute percentage error (*APE*) for a given instance is defined as follows:

$$APE = 100 \times |(x - r)/r|, \tag{1}$$

where in the present context, x is the forecaster's prediction of the exchange rate in question and r is the actual or "realized" rate. *MdAPE* is simply the median value of *APE* over the pertinent collection of judgment cases. Clearly, a forecaster's goal is to minimize *MdAPE*, since in the ideal case, $x = r$.

Table 1 summarizes the median values of *MdAPE* achieved by the professional and amateur forecasters for both horizons, one day and one week. It also shows the ranges of those statistics. As the table indicates, the

Table 1
Medians [ranges] of median absolute percentage error (MdAPE) values for point forecasts by professional and amateur forecasters, for one-day and one-week horizons

Horizon	Forecaster group	
	Professionals	Amateurs
One day	.30%[a***,b**] [.20%, .60%]	.40%[a***] [.20%, .60%]
One week	.90%[b***] [.30%, 1.90%]	1.00% [.50%, 1.60%]

Note. Smaller values of *MdAPE* better.

[a] One-day horizon better than one-week horizon per Wilcoxon signed-ranks test.

[b] Professionals better than amateurs per Mann–Whitney U test.

[**] $p < .01$.

[***] $p < .001$.

professionals' point judgments were not strikingly better than the amateurs' judgments numerically, although the professional-amateur differences were highly significant statistically for both horizons, per Mann–Whitney U tests. The table also shows that, as one might expect intuitively, for both the professionals and the amateurs, one-day-ahead predictions were far more accurate than one-week-ahead forecasts, according to Wilcoxon signed-ranks tests.[1]

These results indicate that it is indeed reasonable to expect that professional FX forecasters will outperform even amateurs who are cognizant of FX financial theory and have extensive first-hand, personal experience in local FX markets. Nevertheless, the data do not suggest that the advantage is a substantial one in absolute terms. The ranges of the *MdAPE* values are noteworthy in this respect, too. Observe in Table 1, for instance, that the weekly point forecasts of the best professional forecaster were far superior to those of the best amateur. Yet, the least successful professional had less accurate weekly forecasts than the least successful amateur. So, practically, a "consumer" of point FX forecasting services (an individual or an organization) in a situation like that which existed in Turkey at the time of the study probably should think hard about the tradeoffs between the costs of expertise and what that expertise might afford in terms of improved decisions. That consumer would also be wise to verify the actual expertise of any source of judgments under consideration—professional or otherwise (including the consumer him-or herself).

On average, both the professional and the amateur participants expected that 20% of their point forecasts would exactly match the actual values of the FX rates they were trying to anticipate, for both one-day and one-week horizons.[2] The median actual percentages of exact matches were 6% for daily forecasts and 2% for weekly forecasts, for both professionals and amateurs. Not surprisingly, the differences between expectations and actual matches were highly significant statistically ($p < .001$), strongly implicating a particular kind of overconfidence.

Another measure of overall point judgment accuracy is the mean squared error (*MSE*), which is defined as follows:

$$MSE = (1/N)\sum(x - r)^2, \tag{2}$$

where the summation extends over all N cases the forecaster considers. As suggested previously, one drawback to *MSE* relative to *MdAPE* is that its numerical values may be difficult to interpret when the quantities of concern lie along scales that have different ranges. But one advantage of *MSE* is that, like similar statistics in regression analyses, it is decomposable into meaningful elements that offer additional insights about various aspects of forecasting performance. In particular, Theil (1966) showed that the following is true

$$MSE = (\bar{x} - \bar{r})^2 + (SD_x - SD_r)^2 + 2(1 - r_{xr})SD_xSD_r. \tag{3}$$

In Eq. (3), in the present context $\bar{x}$ and $\bar{r}$ are the means of the predicted and realized values of the FX rates in question, respectively, and SD_x and SD_r are their standard deviations. And r_{xr} is the correlation between predicted and actual rates. Thus, the "Theil decomposition" indicates that the overall accuracy of point forecasts is a function of three distinct tendencies on the part of the forecaster: (a) the tendency to over- or underpredict actual rates ($\bar{x}$ vs. $\bar{r}$); (b) the tendency to offer forecasts whose variability matches (or fails to match) that of actual FX rates (SD_x vs. SD_r); and (c) the tendency for forecasted rates to covary with actual rates (r_{xr}).

Values of *MdAPE* and *MSE* do not always point toward the same conclusions about the overall accuracy of particular sets of judgments, and such was the case here (e.g., professionals' forecasts yielded statistically significantly better values of *MSE* than amateurs' forecasts only for one-week-ahead forecasts, whereas *MdAPE* had shown better values for both one-day-ahead and one-week-ahead predictions). Since *MdAPE* provides a unit-free measure that trims outliers, while *MSE* remains affected by changes in units as well as extreme errors, such differences are to be expected for the kinds of FX rates studied. For the most part, *MdAPE* is a more suitable measure of overall accuracy when these differences exist. But the elements of point forecasting accuracy implicated by the Theil decomposition of *MSE* can be informative for our explanation question nevertheless.

Table 2 displays the median values of the key statistics distinguished in the Theil decomposition. The table shows that one-day-ahead point forecasts by both professionals and amateurs were better than one-week-ahead forecasts with respect to all three accuracy dimensions. And that display indicates that professional-amateur differences were statistically significant only for one-week horizons. Interestingly, the amateur forecasts were better than the professional predictions with respect to over- and underprediction of FX rates; squared differences of mean forecasts and actual rates were smaller for the amateur participants. It was the other two dimensions that carried the day for the

[1] The sampling distributions for many of the statistics commonly used in analyses of judgment accuracy have not been studied and hence are not well understood. We thus report conservative non-parametric tests even though the results of stricter parametric procedures yielded consistent conclusions.

[2] Strictly speaking, if the participant conceptualized each rate in question as a truly continuous quantity, the sensible expectation was zero. But because of necessary and customary rounding, rates are not, in fact, fully continuous but rather discrete. Thus, as the realization data show, a zero expectation is not fully warranted.

Table 2
Medians of Theil decomposition elements of mean-squared-error (MSE) values for point forecasts by professional and amateur forecasters, for one-day and one-week horizons

Horizon	Forecaster group	
	Professionals	Amateurs
One day	$(\bar{x}-\bar{r})^2$: 5337^{a***} $(SD_x-SD_r)^2$: 8084^{a***} r_{xr}: .999^{a***}	$(\bar{x}-\bar{r})^2$: 3713^{a**} $(SD_x-SD_r)^2$: 7673^{a***} r_{xr}: .999^{a***}
One week	$(\bar{x}-\bar{r})^2$: 190,062 $(SD_x-SD_r)^2$: 66,465^{b**} r_{xr}: .999^{b**}	$(\bar{x}-\bar{r})^2$: 44,315^{c*} $(SD_x-SD_r)^2$: 93,826 r_{xr}: .999

Note. x, point forecast of FX rate; r, actual, realized rate.

[a] One-day horizon better than one-week horizon per Wilcoxon signed-ranks test.

[b] Professionals better than amateurs per Mann–Whitney U test.

[c] Amateurs better than professionals per Mann–Whitney U test.

[*] $p < .05$.

[**] $p < .01$.

[***] $p < .001$.

professionals, the dimensions concerning variability and covariance. Most importantly, the squared differences in the standard deviations of forecasted and actual rates were on average smaller for the professionals than for the amateurs.

The correlations between participants' forecasts and the corresponding actual FX rates (r_{xr}) were especially intriguing. As indicated in Table 2, those correlations were astonishingly high across the board, with the medians reported in the table being within .001 of perfect values of 1.0. In fact, the differences in those median values do not appear until the fourth decimal place. However, the ranges of correlation values were relatively wide (i.e., for professionals, one-day horizon [.9341, .9999], one-week horizon [.9437, .9999]; for amateurs, one-day horizon [.8916, .9999], one-week horizon [.5051, .9999]). In fact, statistically significant differences (in Fisher z-transformed correlation coefficients) revealed (1) higher correlations for daily forecasts as compared to weekly forecasts for both professionals and amateurs ($p < .001$ for both), and (2) professionals showing a better correlation than amateurs for one-week-ahead forecasts ($p = .001$). As discussed by Sheskin (2000), such seemingly inflated r_{xr} values almost certainly resulted from the extremely wide range of data values observed (e.g., from 1.1372 for Swiss franc/US dollar to 90,947 for Turkish lira/British pound at the time this study was conducted). In such circumstances, the correlation coefficient is not an informative indicator of forecasting performance (Armstrong, 2001). Thus, we essentially ignored the correlation values and concentrated on the remaining measures distinguished in the Theil decomposition. And the findings revealed that the tendency to over- or under-predict the realized rates, along with the tendency to offer predictions with a degree of variability matching that of the realized rates, discriminated most the point forecasting performance of professionals and amateurs.

Directional forecasts

Two parallel analyses of directional forecasts were performed. The alternative analyses are distinguished by how a given participant's predictions were encoded, internally vs. externally. Recall that, for a given FX rate, the participant predicted whether the rate would (a) "increase" or else (b) "decrease or remain the same." In "internal coding" as instantiated here, the target event in question was $A =$ "My predicted direction will prove to be correct," and such predictions tend to differ from one instance to the next. This is the kind of coding entailed in most studies of, for instance, overconfidence in general knowledge (e.g., concerning almanac questions like, "Which is farther north, (a) New York or (b) London?"). In contrast, in "external coding" (such as that used in weather forecasting), the focal target event is specified in advance and is the same for every case. In the present research, after the fact and for purposes of analysis, we established a convention such that the target event was $A^* =$ "The rate will increase." For a given case, if the participant predicted a rate increase and reported a "probability-correct judgment" $P'(A) = P \geqslant 50\%$, then we set $P'(A^*) = P$. On the other hand, if the participant predicted a rate *decrease* and specified $P'(A) = Q > 50\%$, we followed normal convention, assumed additivity, and set $P'(A) = 1 - Q$.[3] The reason for performing the dual analyses is that, as noted by Yates (1982), various commonly used accuracy statistics have different (and sometimes problematic) interpretations depending on the kind of coding employed.

Overall accuracy. The probability score (*PS*) is the most popular means for evaluating the accuracy of probability judgments for discrete events, such as whether an exchange rate will increase or that one's predicted directional change will prove correct. Following custom, let f be the probability judgment in question and let d denote an "outcome index," which assumes the value 1 if the target event occurs and 0 otherwise. It sometimes helps intuition to think of d as the probability judgment of a clairvoyant, who reports 100% certainty when the target event in question is going to occur and 0% otherwise. Then we have

$$PS = (f - d)^2. \tag{4}$$

[3] For two-event partitions of sample spaces, such as those here, people's probability judgments for the alternative events generally do sum to 1.0 even though the underlying judgment processes are sometimes not as simple as such reports might suggest (Windschitl, 2000).

Clearly, a forecaster's aim is to minimize PS at 0, for in that instance, judgment is ideal; there is a perfect match between the forecaster's judgment and that of the clairvoyant. And the worst possible performance is indicated when $PS = 1$. The usual measure of accuracy over a given collection of judgment cases is the mean of the probability score for each of the individual cases, $\overline{PS}$, often described as the "Brier (1950) score." It is straightforward to show that the value of $\overline{PS}$ for any particular set of judgments is the same regardless of whether coding is internal or external, and thus a single analysis of overall accuracy is sufficient.

Table 3 presents the median values of $\overline{PS}$ achieved by the professional and amateur participants for both their one-day and one-week forecasts. It also shows the ranges of those statistics. We see that the professionals were significantly more accurate in their directional probabilistic forecasts than the amateurs, especially so in the case of one-week horizons. Nevertheless, as revealed by the ranges, there was considerable overlap in $\overline{PS}$ values for our professional and amateur participants. It is plain to see that the best professional was more accurate than the best amateur. On the other hand, the accuracy of the best amateur was far superior to that of the worst professional and even the *average* professional. All of these indications were consistent with those revealed for point forecasts. But for horizon effects, the indications were markedly different. Recall that one-day-ahead point forecasts were significantly more accurate than one-week-ahead forecasts in that same format. The exact opposite was true for directional probabilistic forecasts. Observe that, for both the professionals and amateurs, the median values of $\overline{PS}$ were distinctly lower (i.e., better) for the weekly forecasts than for the daily forecasts.

The values of $\overline{PS}$ themselves allow direct accuracy comparisons of professionals to amateurs and of one-day horizons to one-week horizons. But it is not immediately obvious what those statistics tell us about how accurate our participants were in "objective" terms. Various standards, such as those shown on the right-hand side of Table 3, thus provide essential reference points. Each of the standards shown there refers to the accuracy level that would have been achieved by a particular kind of fictional "constant judge," one who would have reported the same probability for every case considered (cf., Yates, 1990, pp. 43–44). A "uniform judge" is one who says that all of the alternative specified events are equally likely. (We might imagine a real forecaster adopting the uniform judge's strategy and conceding, "Since I know so little, why don't I just say that all possibilities have the same probability?") So, in the present instance, where there were two alternatives (i.e., that a rate would "increase" vs. "decrease or remain the same"), the reported judgment for the target event would be .50. As is apparent from Eq. (4), a uniform judge (in the two-alternative situation) necessarily always achieves $\overline{PS} = .25$. And as Table 3 shows, on average, the professional and amateur participants, for both one-day and one-week horizons, surpassed the standard set by the uniform judge. But also note that, as indicated by the ranges of $\overline{PS}$ values, not every forecaster met that standard. The least accurate amateur *and* professional participants fell short of this modest benchmark.

A "historical judge" sets a more stringent yardstick. This is an individual who reports that the probability for the target event in a given instance is the relative frequency or "historical base rate" (HBR) with which that event occurred in a particular collection of similar past cases. Recall that, as possible aids, the present participants were provided with graphs and tables displaying past values of each of the FX rates they were asked to forecast. In principle, each participant could have computed the relative frequencies of directional rate changes from those records and then reported for every case, say, $P'(\text{Increase}) = \text{HBR}$, the historical base rate of FX rate increases implicit in the available records. The labor required to do that precisely would, of course, be prohibitive. But roughly estimating the historical base rates from those data was not out of the question. In any

Table 3
Median values [ranges] of mean probability scores $\overline{PS}$, indexing the overall accuracy of probabilistic directional forecasts by professional and amateur forecasters, for one-day and one-week horizons

Horizon	Forecaster group		Standard		
	Professionals	Amateurs	Uniform judge	Historical judge	Base rate judge
One day	$.195^{***b}$ [.120, .288]	.225 [.155, .309]	.250	.202	.182
One week	$.176^{**a,*b}$ [.112, .293]	$.185^{***a}$ [.132, .259]	.250	.212	.202

Note. Smaller values of $\overline{PS}$ better.

[a] One-week horizon better than one-day horizon per Wilcoxon signed-ranks test.

[b] Professionals better than amateurs per Mann–Whitney U test.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

case, we observe in Table 3 that in only one instance—for amateurs making one-day-ahead predictions—did the median participant fail to meet the standard set by the pertinent historical judge. Nevertheless, the least accurate professional and amateur forecasters fell far short of the mark established by the historical judge for both one-day and one-week horizons.

What is commonly called simply a "base rate judge" imposes an even more exacting standard. Suppose that, somehow, it were possible to anticipate the relative frequency or base rate with which a target event actually occurs in the current, given collection of cases, the *sample* base rate (SBR). Then the base rate judge—a kind of "semi-clairvoyant"—would report SBR as the probabilistic forecast for every individual case within that sample. For instance, suppose that (a) the target event is an increase in the focal FX rate, (b) the forecaster will consider 100 cases, and (c), for exactly 55 of those cases, the rate does indeed increase, yielding SBR = .55. Then the base rate judge would report $P'(\text{Increase}) = .55$ for every one of the 100 cases of concern. Now, no one can realistically expect any human forecaster to know in advance what will be the base rate for any particular sample of cases, and hence precise application of the base rate judge's strategy is impossible. Yet, in principle, that semi-clairvoyant strategy could be approximated by, for instance, starting with the historical base rate and adjusting it according to one's hunches as to how conditions during the present sample of cases differ systematically from those during the time when the historical base rate was compiled, i.e., reporting the estimate $\text{SBR}' = \text{HBR} + \Delta$, where Δ is an adjustment per current conditions. As Table 3 indicates, neither the amateur nor the professional participants, on average, outperformed the base rate judge for the one-day-ahead directional probabilistic forecasts. But interestingly, both outperformed the base rate judge for one-week-ahead predictions. Once again, though, there were substantial individual differences such that the least accurate forecasters were greatly outperformed by the pertinent base rate judges.

In terms of the expectations question, the results clearly indicate that it would be most reasonable to anticipate that, in general, professionals' probabilistic directional forecasts would be superior to those of amateurs. It is impossible to directly compare the accuracy of point and probabilistic directional forecasts. Yet, the patterns of the results make it apparent that, speaking to the format question, forecast format makes a big difference. For, although daily point forecasts were more accurate than weekly point forecasts, the opposite was true for directional probabilistic predictions. Analyses of various dimensions of overall probability judgment accuracy, to which we turn next, allow for insights about the explanation question.

Dimensional accuracy—the internal coding perspective. As described below (cf., Yates, 1982), there exist several schemes for decomposing $\overline{PS}$ into informative components. But when coding is internal, the component that is most cleanly interpreted is one called "bias" and is indexed by the following statistic

$$Bias = \bar{f} - \bar{d}, \tag{5}$$

where, using the notation introduced in Eq. (4), $\bar{f}$ is the mean probability judgment reported for the event in question and $\bar{d}$ is the mean of the outcome index. Since d is 1 when the pertinent event occurs and 0 when it does not, it is clear that $\bar{d}$ is the same as the proportion of times that the target event actually occurs. Thus, the bias statistic is an indicator of the extent to which the forecaster, in effect, over- or underpredicts the target event. Recall that, for internal coding in the present instance, the target event was, from the forecaster's perspective, $A =$ "My predicted direction (for an FX rate change) will prove to be correct." Therefore, it is reasonable to interpret the bias statistic as a measure of case-level, in-the-moment overconfidence when it is positive and an index of underconfidence when it is negative. (The "case-level, in-the-moment" qualifier is explained later.)

As shown in Table 4, the participants were generally overconfident, but with one exception—when the professionals were making one-day-ahead predictions. In this latter instance, there was a tendency for forecasts to be slightly underconfident. The amateur participants were typically more overconfident than their professional counterparts, although the difference was statistically significant only in the case of daily forecasts. (Experience

Table 4
Median values [ranges] of the bias statistic ($\bar{f} - \bar{d}$) for internally coded probability judgments, indexing case-level, in-the-moment over- or underconfidence in probabilistic directional forecasts by professional and amateur forecasters, for one-day and one-week horizons

Horizon	Forecaster group	
	Professionals	Amateurs
One day	−.023[b*,c**] [−.188, .175]	.051 [−.122, .234]
One week	.031 [−.235, .290]	.040[a*] [−.113, .181]

Note. Internal coding: target event $A =$ "My predicted direction will prove to be correct"; f, probabilistic judgment; $d = 1$ if actually correct, $d = 0$ otherwise, $Bias = 0$ ideally, indicating neither overconfidence nor underconfidence.

[a] One-week-ahead forecasts less overconfident that one-day-ahead forecasts per Wilcoxon signed-ranks test.

[b] One-day-ahead forecasts less overconfident that one-week-ahead forecasts per Wilcoxon signed-ranks test.

[c] Professionals less overconfident than amateurs per Mann–Whitney U test.

[*] $p < .05$.

[**] $p < .01$.

breeds caution?) Also, whereas the professionals were more overconfident for their weekly forecasts relative to their daily forecasts (which were actually underconfident), the opposite was true for the amateurs.

The results summarized in Table 5 offer a different perspective on the notions of over- and underconfidence. Each cell in the table first presents the median value of the participants' personally articulated expected percentages of correct directional forecasts, expressed as proportions to permit easier comparisons to Table 4. The table also shows the corresponding actual or realized percentages of correct directional predictions as well as the differences between the expected and realized percentages. In parallel to the previous bias statistics, such differences measure a second, "aggregate-level, anticipation" variety of overconfidence (when positive) or underconfidence (when negative). The statistics presented earlier in Table 4 rested on the judgments that participants rendered at the very moment when each individual FX case was considered (e.g., "The US dollar/Japanese yen rate will increase next week, and I'm 80% sure that that will happen"). On the other hand, those in Table 5 derived from the participant's aggregate percentage-correct estimate in advance of considering any concrete cases (e.g., "About 60% of my directional predictions will prove correct"). Table 5 shows that, in terms of the accuracy of categorical directional forecasts per se, the predictions of the professionals were significantly better than those of the amateurs, for both the one-week horizon and especially the one-day horizon. Recall from Table 4 that participants were typically overconfident in the individual, case-level directional forecasts they offered at the times when they actually deliberated those cases. In marked contrast, the Table 5 comparisons, between expected and realized percentages, reveal pervasive *underconfidence* in aggregate-level, personally expressed expectations, for the professional participants in particular.

Table 5
Median values of personally expected and realized percentages correct (expressed as proportions) for probabilistic directional forecasts by professional and amateur forecasters, for one-day and one-week horizons

Horizon	Forecaster group	
	Professionals	Amateurs
One day	Expected: .600 Realized: .733[a***,b**] Difference: −.133	Expected: .600 Realized: .640 Difference: −.040
One week	Expected: .500 Realized:.740[a*,b***] Difference: −.240	Expected: .600 Realized: .700[b***] Difference: −.100

[a] Professionals better than amateurs per Mann–Whitney *U* test.

[b] Realized% correct higher than expected% correct, implying aggregate-level, anticipation underconfidence per Wilcoxon signed-ranks test.

[*] $p < .05$.

[**] $p < .01$.

[***] $p < .001$.

The discrepancy between the two different kinds of over- and underconfidence shown here—"case-level, in-the-moment" vs. "aggregate-level, anticipation"—is reminiscent of similar differences revealed in previous research. A good example is the study reported by Lee et al. (1995). These investigators required participants to perform the type of task common in most studies of general knowledge overconfidence (e.g., "Potatoes grow better in (a) warm or (b) cool weather? Now, how sure are you (50–100%) that your chosen answer is actually correct?"). The participants also performed a peer comparison task in which they estimated the percentage of their peers to whom they were superior in various domains (e.g., writing skills). Lee et al. found virtually no correlation between the two different varieties of over- and underconfidence they observed, and they concluded that those constructs rest on qualitatively different mechanisms. The same conclusion is reasonable here. Thus, the overconfidence revealed for individual probabilistic directional FX forecasts likely arise from factors such as the forecaster failing to bring to mind, in the moment, specific arguments that disagree with the forecaster's expected rate change direction. In contrast, when the forecaster must estimate how many of his or her direction predictions will prove correct in a collection of 50, the forecaster plausibly draws on recollections of what happened in generically similar situations in the past (cf., Gigerenzer, Hoffrage, & Kleinbölting, 1991).

It also seems reasonable that the differences in the indications of over- and underconfidence evident in Tables 4 and 5 are at least partly due to the elicitation of 0–100% probability judgments for individual forecasts but 0–50 frequency estimates for personal expectations. Price (1998), in a general knowledge study, obtained results completely parallel to the ones reported here. In Price's instantiation of the frequency format, for any given question, he asked the participant: "Out of 100 questions for which you felt this certain of the answer, how many would you answer correctly?" This format resulted in much lower overconfidence than the usual probabilistic judgment format. Price, too, argued that it is likely that the alternative formats induce different kinds of reasoning that are differentially susceptible to overconfidence.

Another aspect of the results presented in Table 5 is noteworthy, too, one that bears directly on the format question. Observe that the realized percentages of correct directional predictions were quite good, ranging from 64 to 74%. But recall the earlier Theil decomposition analysis of the participants' point forecasts. In particular, consider once again the correlations (r_{xr}) between the participants' forecasts and the actual values of the pertinent FX rates, as displayed in Table 2. Those correlation coefficients were astronomical, on average, nearly 1.0. Such correlations should translate to correct directional prediction percentages substantially higher than the ones that participants actually achieved (cf., Kendall, 1948). This discrepancy suggests that participants necessarily possessed the knowledge required to support the superb directional forecasts implicit in the point forecasts they reported when given an instruction equivalent to, "What do you think that rate will be?" Yet, the demand to exercise their abilities explicitly, in response to a request amounting to, "Do you think that rate will increase?" somehow caused that knowledge to become misdirected.

Dimensional accuracy—the external coding perspective. One well-known scheme for analyzing the accuracy of probability judgments derives from the Murphy (1973) decomposition of the mean probability score, whose formal expression is given in Appendix B. Fig. 1 (cf., Yates, 1994) provides a heuristic way to think about what the Murphy decomposition can reveal about how the present participants plausibly arrived at their particular levels of probabilistic forecasting accuracy. In the schematic shown there, as before, $\bar{d}$ is the base rate for the target event. *CI* is the "calibration index," which is defined in Appendix B, and *DI* is the "discrimination index," whose analytic definition is also given in Appendix B.

The first accuracy dimension distinguished in the Murphy decomposition is a particular kind of task difficulty. A target event (when there are two alternatives) whose base rate ($\bar{d}$ in the present notation) is close to 50% is inherently harder to predict than one whose base rate is close to 0% or 100%; there is less fundamental uncertainty, in the intuitive and information theory senses of the term. As the expression for the decomposition shows (i.e., outcome index variance $= \mathrm{Var}(d) = \bar{d}(1-\bar{d})$), the greater is this uncontrollable uncertainty (i.e., the closer is $\bar{d}$ to .5), then the worse (i.e., higher) is $\overline{PS}$, through no fault or virtue of the forecaster when the target event is externally coded, such as when $A^* =$ "The rate will increase," as in the present instance. Therefore, legitimate accuracy comparisons of one group of forecasters to another (e.g., professionals to amateurs) or forecasts in one context to those in another (e.g., for one-day-ahead vs. one-week-ahead predictions) should focus on the remaining, controllable accuracy dimensions.

Probability judgments are said to be "well calibrated" to the extent that the numerical values attached to those judgments match the relative frequencies with which the target event actually occurs. Thus, suppose there are 50 occasions on each of which a weather forecaster says there is a 40% chance that precipitation will occur within the next 12 hours. Then, if that forecaster's judgments are perfectly calibrated, precipitation will in fact be observed on exactly 20—that is, 40%—of those occasions. Clearly, calibration is a judgment accuracy dimension under a forecaster's control. It is measured by the calibration index, *CI*. We see in Table 6 that calibration was significantly better for the professional participants, but only when the horizon was one day away.

Discrimination is the other accuracy dimension distinguished in the Murphy decomposition of $\overline{PS}$. As described in Fig. 1, discrimination has nothing to do with the specific numbers a forecaster uses in expressing his or her opinions about what is going to happen in the future. Instead, it concerns the non-numerical association between the qualitatively *different* things the

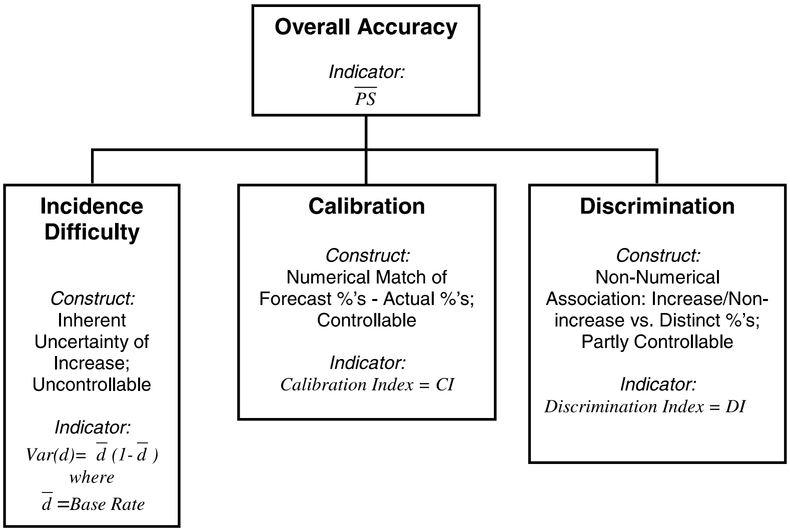
Overall Accuracy
Indicator:
$\overline{PS}$
Incidence Difficulty
Construct:
Inherent Uncertainty of Increase; Uncontrollable
Indicator:
$Var(d)= \bar{d}\,(1-\bar{d}\,)$
where
$\bar{d}$ =Base Rate
Calibration
Construct:
Numerical Match of Forecast %'s - Actual %'s; Controllable
Indicator:
Calibration Index = CI
Discrimination
Construct:
Non-Numerical Association: Increase/Non-increase vs. Distinct %'s; Partly Controllable
Indicator:
Discrimination Index = DI

Fig. 1. Schematic representation of overall probabilistic judgment accuracy and its elements as discerned in the Murphy (1973) decomposition of the mean probability score.

forecaster says, on the one hand, and what actually occurs in the future, on the other. The discrimination index, *DI*, measures the degree of discrimination a forecaster achieves. As noted in Appendix B, the fact that discrimination is about non-numerical association is reflected in the formal connection between *DI* and the Pearson χ^2 statistic commonly used in contingency table analyses. Table 6 shows that the discrimination exhibited by the professional participants was superior to that of the amateurs for the one-day horizons and especially the one-week horizons. Good discrimination requires two main things. The first is access to information that is reliably associated with the target event—valid "cues" in the common parlance of the judgment literature. In this context, this would include facts that, for whatever reason, tend to be correlated with FX rate changes. The second prerequisite for good discrimination is knowledge and skill at actually using such predictive cues, including mere attention to those cues. Accessibility is often out of the forecaster's control. Therefore, one plausible reason that the professional forecasters might have exhibited superior discrimination is that their affiliated institutions or companies made readily available to them useful facts that are less available to amateurs. But the professionals might also have, over time, acquired routines for more appropriately interpreting facts accessible to both professionals and amateurs.

The covariance decomposition of $\overline{PS}$ provides another, more fine-grained means for analyzing overall probability judgment accuracy into usefully distinguished components (Yates, 1982). Fig. 2 sketches the various accuracy dimensions distinguished in the covariance decomposition. As is immediately apparent, the first component, concerning incidence difficulty, is shared with the Murphy decomposition. The others, however, are different. The equations for those elements as well as for the decomposition itself are presented in Appendix B.

Table 6
Median values [ranges] of calibration (CI) and discrimination (DI) indexes for externally coded probabilistic directional forecasts by professional and amateur forecasters, for one-day and one-week horizons

Horizon	Forecaster group	
	Professionals	Amateurs
One day	*CI*: .047[b***]	*CI*: .072
	[.017, .134]	[.012, .163]
	DI: .022[b*]	*DI*: .019
	[.004, .090]	[.001, .086]
One week	*CI*: .044	*CI*: .035[a***]
	[.002, .184]	[.003, .119]
	DI: .054[a***,b**]	*DI*: .031[a***]
	[.004, .311]	[.003, .094]

Note. External coding: target event $A^* =$ "the rate will increase"; smaller values of *CI* better; larger values of *DI* better.

[a] One-week horizon better than one-day horizon per Wilcoxon signed-ranks test.

[b] Professionals better than amateurs per Mann–Whitney U test.

[*] $p < .05$.

[**] $p < .01$.

[***] $p < .001$.

Formally, the bias component is identical to the one discussed before. Here, however, the target event is different. Recall that, in the previous discussion of internal coding, the target was $A =$ "My predicted direction will prove to be correct." In the present external coding context, it is $A^* =$ "The rate will increase." That is why the appropriate interpretation of the *Bias* statistic in this context is that it describes the degree to which a forecaster tends to overpredict FX rate increases (when the statistic is positive) or underpredict such increases (when it is negative). *Bias* reflects a coarse variety of calibration (sometimes called "calibration in the large") and, as suggested in Fig. 2, is largely controllable. As indicated in Table 7, both the professionals and amateurs tended to underpredict FX increases, and significantly more so in the case of one-day-ahead forecasts. Although the bias within the daily forecasts was statistically significantly better for the professionals than for the amateurs, for the most part, bias did not sharply distinguish the forecasting performance of professionals and amateurs.

Suppose that, for a given forecaster, two distributions of probabilistic judgments $f = P'$ (Increase) were constructed, where "Increase" means that the FX rates in question will increase. The first distribution consists of judgments rendered in cases where the pertinent rates eventually did indeed increase. The second distribution is comprised of similar judgments in the opposite kinds of cases, when the rates of concern did *not* increase. For a clairvoyant, all the judgments in the former "conditional" distribution would have been $f = 1.0$, whereas every one in the latter would have been $f = 0$. Suppose that we denote the mean of the judgment distribution conditional on an actual rate increase by $\overline{f_1}$ and that for the distribution conditional on an actual non-increase by $\overline{f_0}$. Then for a clairvoyant, the difference in these means, $\overline{f_1} - \overline{f_0}$, is necessarily 1.0. A real, human forecaster can only approximate this ideal. To the extent that the difference—called the "slope"—does indeed approach the ideal of 1.0, in effect, the forecaster has approached maximum separation of the conditional distributions, the same as a clairvoyant. In order to do this, the forecaster must have access and pay attention to cues that have a reliable relationship to FX rate changes. The forecaster must also be skilled at attaching appropriate numerical labels to forecasts. Thus, the separation construct is a particular combination of the discrimination and calibration constructs distinguished in the Murphy decomposition. As shown in Table 7, separation was clearly the primary means by which the professionals in the present study outshone the amateurs, for both the one-day and one-week horizons.

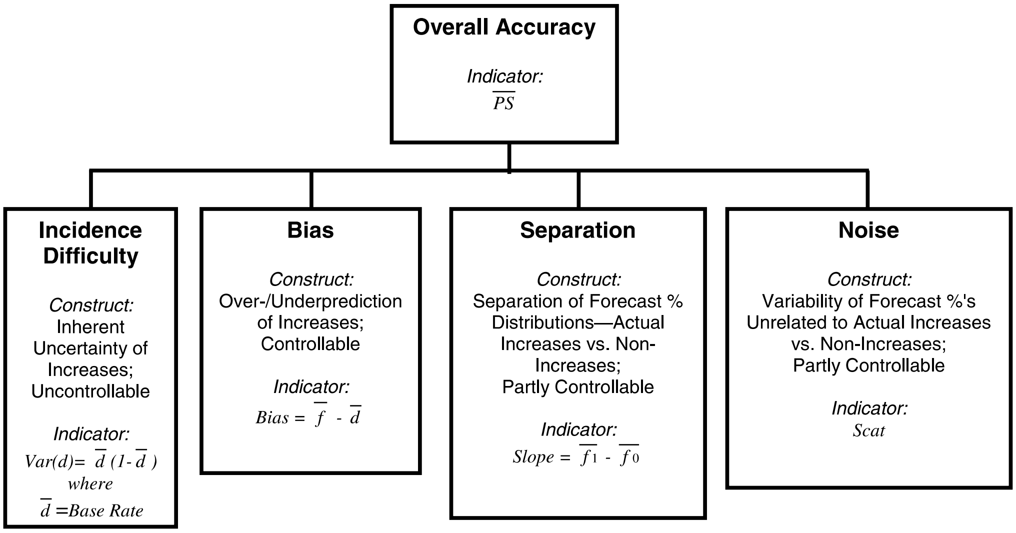
Overall Accuracy
Indicator:
$\overline{PS}$
Incidence Difficulty
Construct:
Inherent Uncertainty of Increases; Uncontrollable
Indicator:
$Var(d)= \bar{d}\,(1-\bar{d})$
where
$\bar{d}$ =Base Rate
Bias
Construct:
Over-/Underprediction of Increases; Controllable
Indicator:
$Bias = \bar{f} - \bar{d}$
Separation
Construct:
Separation of Forecast % Distributions—Actual Increases vs. Non-Increases; Partly Controllable
Indicator:
$Slope = \bar{f}_1 - \bar{f}_0$
Noise
Construct:
Variability of Forecast %'s Unrelated to Actual Increases vs. Non-Increases; Partly Controllable
Indicator:
Scat

Fig. 2. Schematic representation of overall probabilistic judgment accuracy and its elements as discerned in the covariance decomposition of the mean probability score.

Table 7
Median values [ranges] of *Bias*, *Slope*, and *Scat* statistics for externally coded probabilistic directional forecasts by professional and amateur forecasters, for one-day and one-week horizons

Horizon	Forecaster group	
	Professionals	Amateurs
One day	*Bias*: −.129[b*]	*Bias*: −.163
	[−.258, .006]	[−.246, .005]
	Slope: .105[b***]	*Slope*: .035
	[−.048, .353]	[−.151, .250]
	Scat: .032	*Scat*: .030
	[.006, .114]	[.007, .096]
One week	*Bias*: −.036[a***]	*Bias*: −.054[a***]
	[−.175, .110]	[−.162, .085]
	Slope: .195[a***,b**]	*Slope*: .128[a***]
	[−.073, .447]	[−.021, .356]
	Scat: .034	*Scat*: .032
	[.007, .135]	[.008, .107]

Note. External coding: target event $A^* =$ "The rate will increase"; $Bias = 0$ ideally; larger values of *Slope* better; smaller values of *Scat* better.

[a] One-week horizon better than one-day horizon per Wilcoxon signed-ranks test.

[b] Professionals better than amateurs per Mann–Whitney *U* test.

[*] $p < .05$.

[**] $p < .01$.

[***] $p < .001$.

Table 7 also makes it clear that separation was what made weekly directional probabilistic forecasts so much better than daily forecasts (see Table 3 for overall accuracy comparisons), even though the outcome index variance was higher for the weekly as compared to the daily realized values.

"Noise" is the final accuracy dimension distinguished in the covariance decomposition of $\overline{PS}$. Analogous to error variance in the analysis of variance, the noise construct refers to variability in an individual's probabilistic forecasts that is unrelated to "the truth," that is, whether the FX rates in question in actuality increase or fail to increase. Noise can arise in either or both of two ways. On one hand, it can be a manifestation of pure unreliability in the forecaster's judgment processes. Such unreliability would be revealed in inconsistencies between the predictions the forecaster offers on two different occasions when presented with identical facts. It is hard to imagine perfect replications of real-life FX conditions, but conceptually, the idea is the same as that of parallel forms of a psychological test. Unreliability would be manifested to the extent that the test–retest reliability coefficient, r_{tt}, was less than 1.0. In principle, the pure unreliability basis for noise is controllable. One approach to doing this is to replace the human forecaster by a bootstrapping model of that person (cf., Dawes, 1979). Noise can also occur even when a forecaster is perfectly reliable in the test–retest reliability sense. In particular, suppose that the cues or information the forecaster uses to form his or her forecasts are themselves only weakly associated with the target event. Then this guarantees that the forecaster's judgments cannot be strongly related to the target event either.

As shown in Appendix B, in the covariance decomposition of $\overline{PS}$, noise is indexed by a weighted mean of the variances of the distributions of probabilistic forecasts conditional on the target event occurring and not occurring, respectively, a statistic called "scatter" and denoted by *Scat*. We see in Table 7 that the values of *Scat* were virtually identical for the professional and amateur forecasters and for the one-day and one-week horizons. The professional-amateur null effect is especially noteworthy. The reason is that, in the studies of stock price and earnings forecasting by Yaniv, Yates, and Smith (1991), noise was the key basis for distinction between experienced and novice forecasters. In those studies, the more experienced forecasters were *less* accurate than their less experienced counterparts because their judgments were noisier. We address this most interesting difference in results in General discussion.

Interval forecasts

The "inclusion rate" for a forecaster's interval forecasts is the proportion of times that the intervals articulated by the forecaster actually capture the true values of the quantities in question. From this perspective, overall interval forecasting accuracy is good to the extent that the inclusion rate is close to the ideal of 1.0. Table 8 shows that, in these terms, the interval forecasts of both the professional and amateur participants were significantly more accurate for the one-day horizon than for the one-week horizon. This table also indicates that there were no reliable professional-vs.-amateur differences in the accuracy of interval predictions for either horizon. Thus, in terms of our expectations question, there is no evidence that we should expect professional forecasters to outperform the kinds of informed amateurs employed in the present study, when it comes to interval forecasts. Recall that, as in the present instance, one-day-ahead point forecasts were more accurate than one-week-ahead forecasts of that type. In contrast, the opposite pattern was observed for directional forecasts. Moreover, the lack of professional-amateur distinctions in the case of interval forecast accuracy is different from what was seen for both point and directional forecasts. Hence, once again, there are indications that, at minimum, one answer to our format question is that formats definitely make a difference in the way professionals exhibit their forecasting expertise.

Table 8
Median values [ranges] of actual and expected inclusion rates, *Log Widths*, and *Normalized Errors* for 90% credible intervals reported by professional and amateur forecasters, for one-day and one-week horizons

Horizon	Forecaster group	
	Professionals	Amateurs
One day	*Actual inclusion rate*: .540[a]** [.160, .940] *Expected inclusion rate*: .600 *Log Width*: .682[a]*** [−.367, 2.996] *Normalized Error*: .520[a]* [.071, 2.550]	*Actual inclusion rate*: .560[a]*** [.122, .880] *Expected inclusion rate*: .600 *Log Width*: 1.066[a]*** [−1.613, 2.303] *Normalized Error:* .445[a]*** [.143, 5.000]
One week	*Actual inclusion rate*: .440 [.140, 1.000] *Expected inclusion rate*: .500 *Log Width*: 1.498 [−3.404, 4.700] *Normalized Error*: .664 [.055, 13.033]	*Actual inclusion rate*: .400 [.040, .760] *Expected inclusion rate*: .600[b]*** *Log Width*: 1.431 [−3.817, 2.708] *Normalized Error*: .704 [.278, 44.611]

Note. Inclusion rate = .90 ideally; smaller values of *Log Width* and *Normalized Error* better.

[a] One-day horizon better than one-week horizon per Wilcoxon signed-ranks test.

[b] Expected rate higher than realized rate per Wilcoxon signed-ranks test.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

A trivial way that an FX forecaster can achieve perfect, 100% inclusion rates is to always report the entire non-negative real line $[0, \infty)$ as a forecast interval. The problem with this strategy, of course, is that such intervals are completely uninformative and therefore useless. This is the motivation for analytic schemes that examine specific elements of overall interval forecasting accuracy, schemes that also permit conclusions about our explanation question.

Most discussions about the accuracy of probabilistic interval forecasts focus on what is sometimes called "distribution calibration" (Yates, 1990, pp. 69–71). Thus, as in the present research, 90% credible intervals would exhibit perfect distribution calibration if their inclusion rates were exactly 90%. As Table 8 makes clear, the distribution calibration of forecasts by both the professional and amateur participants and for both the one-day and one-week horizons was very weak; the inclusion rates fell far short of .90. The particular form of distribution miscalibration observed here has been repeatedly reported since Alpert and Raiffa's (1982) work with general knowledge questions, extending even to judges predicting their own task performance (Connolly & Dean, 1997). The common interpretation of this finding is that it reflects interval overconfidence. The rationale for the terminology in the current context would be that, for instance, although the typical professional participant was 90% sure that each of his or her one-day-ahead 90% credible intervals would capture the FX rate in question, only about 54% of those intervals actually did that. That is, the forecaster was overly certain that the intervals he or she constructed would capture the target rates.

The present findings offer a somewhat different perspective on distribution overconfidence. Table 8 shows the medians of the participants' explicitly articulated expected inclusion rates, based on their responses to the question (see Method), "In how many cases (out of 50) do you expect the realized value to fall within your prediction interval?" Despite the clear and understood specification that, for each interval, "you should be 90% confident" (see Appendix A), on average, expected inclusion rates were much lower than .90. In fact, as shown in Table 8, only in the case of one-week-ahead predictions by amateurs were forecasts overconfident in the sense that actual inclusion rates were statistically significantly smaller than self-reported expected rates. This result agrees with the earlier findings for directional forecasts. Recall that an in-the-moment forecast for an individual FX rate entailed two stages. The participant first predicted whether the rate would "increase or not." The participant then stated a 50–100% probability judgment that that directional prediction would prove to be correct. These forecasts were generally overconfident in that the average probability judgments outstripped the actual proportions of correction directional predictions (Table 4). The participants were asked to explicitly predict their proportions of correct directional predictions, too. Those estimates turned out being consistently smaller than the actual proportions, that is, *under* confident.

In the very best of circumstances, a forecast interval would be degenerate in a special way, one that captures the true value of the quantity in question but whose maximum and minimum values were the same. That is, it would be **a** perfectly accurate point forecast. Short of that ideal, a forecast interval is good to the degree that it captures the realized value and is also narrow. This is the intuition behind the measures of interval forecast accuracy proposed by Yaniv and Foster (1995, 1997). The first measure is termed "informativeness" and is indexed by the natural logarithm of a forecast interval's length:

$$Log\ Width = \ln(U - L), \tag{6}$$

where U is the upper bound of the interval and L is its lower bound. Clearly, all else being the same, one forecast interval is better than another if its width is smaller. That is, the forecaster's aim is to minimize *Log Width*. The second measure of Yaniv and Foster integrates both point and interval forecasts:

$$Normalized\ Error = (|r - x|)/(U - L), \tag{7}$$

where, as before, x is the forecaster's point forecast for a particular FX rate and r is the actual or realized rate. Again, it is apparent that minimization of *Normalized Error* should be the forecaster's goal. Observe in Table 8 that the results for informativeness and normalized error mirror those for inclusion rates. That is, on each of these dimensions, the participants' interval forecasts were better for one-day-ahead predictions than for one-week-ahead predictions. Moreover, there were no statistically significant differences between the professionals and amateurs.

General discussion

Here we will summarize what the present data imply for the expectations, format, and explanation questions that were posed at the outset. We will also highlight key aspects of those questions that remain unresolved.

Expectations

The present findings demonstrate that FX rates are not unpredictable. Participants were, on average, able to surpass accuracy benchmarks of varying degrees of stringency. Thus, we should not expect reasonably well-informed FX forecasters—professional or otherwise—to flounder hopelessly. The results also indicate that we should anticipate that, more often than not, professionals are capable of rendering more accurate FX forecasts than sophisticated amateurs. Importantly, this professional advantage apparently does not exist in some domains, such as that of stock prices and earnings (Staël von Holstein, 1972; Yates et al., 1991). To be sure, there are important qualifications to the general conclusion that professionals enjoy an accuracy advantage over amateurs when predicting FX rates. For one thing, there is likely to be considerable overlap in the distributions of forecasting competencies of populations of professional and amateur FX forecasters. That is, many amateurs will outperform many professionals. In practical terms, this fact reinforces a maxim that consumers of forecasting services should respect generally anyway: Do not simply assume that a prospective source of judgment expertise (including oneself) is in fact expert, on the basis of credentials, reputation, or anything else. Instead, insist on empirical verification of the source's ability to provide forecasts that are reliably predictive of the truth, e.g., using the kinds of accuracy measures illustrated here.

Format

Another qualification to the expectation of superior accuracy on the part of professionals concerns how forecasters are asked for their FX rate predictions. Although the accuracy of professionals was generally better than that of amateurs for point and probabilistic directional forecasts, that superiority washed out for interval forecasts. It is difficult or impossible to directly compare the accuracy of predictions reported in different formats. Nevertheless, the patterns in the present data strongly suggest that formats can greatly affect how accurately forecasters make their predictions and, therefore, the processes by which they arrive at those predictions. One such pattern was the reversal in the relative accuracy of one-day-ahead and one-week-ahead point and interval forecasts, on the one hand, and probabilistic directional forecasts, on the other. Another was the marked difference in the accuracy of directional predictions implicit in participants' point forecasts and those explicit in their probabilistic directional forecasts. A stiff challenge for future studies is determining exactly why these format effects occur. A plausible initial hypothesis is that mere familiarity and experience with formats play a role. In everyday practice, forecasters are far more accustomed to making deterministic and point predictions rather than probabilistic and interval ones.

A further, especially noteworthy finding can be viewed in terms of formats, too, although in a different sense of the term. There were strong indications of overconfidence in the case-level, in-the-moment predictions participants reported in the directional and interval formats. Yet, when the cumulative accuracy levels that participants actually achieved were compared to the levels they had originally said that they expected to achieve, if any bias was evident, it was underconfidence rather than overconfidence. That is, whether we should expect overconfidence in FX forecasts (or perhaps any other judgments) depends on when and how people are asked to express their confidence.

Explanation

The data in hand point toward at least some accounts for the observed effects. But it is clear that the most difficult challenges ahead concern explanations and their implications for practical matters such as forecaster skill development. One finding that begs for explanation is the superiority of the present professional FX forecasters over their amateur competitors when previous work in the context of stock price and earnings predictions yielded conflicting results (e.g., Önkal & Muradoglu, 1994; Yates et al., 1991). The $\overline{PS}$ decomposition analyses reported here suggests a proposal that should be subjected to further, rigorous tests. Earlier research found that experienced forecasters were less accurate than novices in their probabilistic forecasts for stock prices and earnings, at least for the particular forecasting formats used in that research. This seemed to occur because the experienced forecasters relied on information they *thought* was associated with prices and earnings but which really was not, thereby yielding greater noise. Nothing like this was evident in the present FX data. Instead, the professionals appeared to achieve greater accuracy than the amateurs via superior discrimination, calibration, and slope. Excellence with respect to these accuracy dimensions rests on factors such as reliance on cues that are truly predictive and on memory-supported matching of probability reports and relative frequencies. A reasonable hypothesis for the observed differences between the forecasting of the stock prices and earnings, on the one hand, and FX rates, on the other, is the following: Prices and earnings for the myriad firms on the market are affected by a vast array of forces, many of which may be specific to the individual firms and inaccessible to outsiders, including forecasters. Nevertheless, it is tempting and easy for stock forecasters to assemble plausible—but often difficult-to-test—theories for how to make good predictions. In contrast, FX forecasters must concern themselves with a more limited and manageable number of highly interdependent FX rates whose dynamics may be comparatively more traceable and learnable. Additionally, there is reason to suspect that the choice of forecast elicitation format is an important determinant of the extent to which the professionals actually display different dimensions of their forecasting expertise, regardless of the contextual contingencies of their respective financial markets (Önkal & Muradoglu, 1996; Önkal-Atay, Thomson, & Pollock, 2002).

Appendix A. Instructions and form

Please note that the instructions and sample form given below are for daily forecasts only. Instructions and forms for the weekly forecasts were identical except that '11 a.m.' specifications were replaced with 'Monday opening' specifications.

Instructions

In this part of our study, we request that you make forecasts for the DAILY values of various FX rates. We expect you to make forecasts for values that will be realized at *11:00 a.m. tomorrow*. For each of the rates in question, we request that you state your forecasts using three formats:

*POINT FORECASTS: Please write down the value that you think will be realized at 11 a.m. tomorrow.

*INTERVAL FORECASTS: Please write down the lowest and the highest value that this rate could take on with 90% confidence. In other words, you should be 90% confident that tomorrow's 11 a.m. value will fall between these two values (i.e., will fall within this interval).

*DIRECTIONAL FORECASTS: First, please indicate whether the value that will be realized at 11 a.m. tomorrow will increase or not, compared to the value observed today at 11 a.m. After predicting this direction of change, please indicate the probability that your forecast will indeed occur. This will be your subjective probability that the realized change will actually fall in the direction you predicted. Please note that, since you will predict a direction for change first, this probability will have to be between 50 and 100%. If you specify 100% as your probability, this would mean that you are absolutely certain (with no doubts whatsoever) that the realized change will fall in the direction you predicted. If you specify 50% as your probability, this would mean that you believe there's an equal chance for the realized change to fall in your predicted direction (indicating your belief in an equal chance for an 'increase' vs. 'stay the same or decrease' in the rate). Of course, you can give any probability between 50 and 100%. Please keep in mind that increasing percentages reflect stronger beliefs in predicted direction actually occurring.

Please note once again that you should never use a probability of less than 50%, since this would mean you should of have indicated the other direction as your predicted direction of change. For example, if you predicted an increase and then gave a 30% probability, this would indicate that you believe there is a 70% chance of no-increase (i.e., stay the same or decrease), in which case you should of have indicated the 'stay the same or decrease' direction as your prediction, assigning a value of 70% to it.

Illustrative form

YEN/TL (DAILY FORECASTS)

*The value I think will be realized at 11 a.m. tomorrow: ___________

*I am 90% confident that the value of this FX rate that will be realized at 11 a.m. tomorrow will be between ___________ and ______________

*When compared to today's 11 a.m. value, tomorrow's 11 a.m. value will:

A. Increase

B. Stay the same or decrease

Your forecast (A or B) : ___________

Probability that your forecast will indeed occur (i.e., probability that the daily change will actually fall in the direction you predicted) (BETWEEN 50% AND 100%) :

____

Appendix B. Formulas for probability judgment accuracy dimension indicators

Murphy decomposition

Calibration index. Suppose that all probabilistic forecasts f for the target event (e.g., an FX rate increase) are rounded to particular categories, e.g., the nearest tenth, .0, 0, .1,...,1.0, which can be represented as f_k, for $k = 1, \ldots, K$, where K is the total number of categories. Then the calibration index (CI) is given by (Murphy, 1973; Sanders, 1963):

$$CI = (1/N)\sum_k N_k(f_k - \overline{d_k})^2. \tag{B.1}$$

In this expression, $N = N_1 + N_2 + \cdots + N_k$ is the total number of forecasts, where N_k is the number of forecasts falling into the kth category, i.e., taking on value f_k. And $d_{k,i}$ is the outcome index for the judgment in specific, individual case i, assuming the value 1 when the target event occurs in that instance and 0 otherwise. Thus, the mean of $d_{k,i}$ over all N_k cases when f_k is reported is $\overline{d_k}$ and is also the proportion of times the target actually occurred out of those cases. CI clearly measures the extent to which the numerical value for a particular forecast category matches the relative frequency with which the target event actually occurs when that forecast is rendered.

Discrimination index. Using the same notation as above, the expression for the discrimination index (DI) is as follows (Murphy, 1973):

$$DI = (1/N)\sum_k N_k(\overline{d_k} - \overline{d})^2. \tag{B.2}$$

Here, $\overline{d}$ is the mean of the outcome index over all N cases, the overall base rate or proportion of times the target event has occurred (e.g., the incidence of increases for the FX rates under consideration). Note that f_k plays no role in DI. That is, DI is unaffected by the numerical character of the reported forecasts. Instead, it reflects the degree to which the forecaster tends to report *different* judgments on the occasions when the target event occurs ($d_{k,i} = 1$) as opposed to those when it does not ($d_{k,i} = 0$), regardless of the numbers attached to those judgments. In effect, DI is a measure of category association akin to the Pearson χ^2 statistic (Yaniv et al., 1991).

Murphy decomposition. Murphy (1973) showed that the following relation holds

$$\overline{PS} = \overline{d}(1 - \overline{d}) + CI - DI. \tag{B.3}$$

Covariance decomposition

Bias. Suppose that $\overline{f}$ is the mean probability judgment reported for the target event and $\overline{d}$ is, again, the overall mean of the outcome index. Then the *bias* statistic is given by

$$Bias = \overline{f} - \overline{d}. \tag{B.4}$$

Slope. Let $\overline{f_1}$ be the mean probability judgment reported for the target event on those particular occasions when it ultimately turns out that that event actually occurs, and let $\overline{f_0}$ be the corresponding average for the remaining instances when that event does not in fact occur. The *slope* statistic for the forecaster's judgments is then represented as

$$Slope = \overline{f_1} - \overline{f_0}. \tag{B.5}$$

Scatter. Suppose that Var(f_1) is the variance of the forecaster's probabilistic forecasts for the target event for the N_1 cases in which the target event actually happens and that Var(f_0) is the corresponding statistic for the remaining N_0 instances when the target does not occur. Then the *scatter* statistic (*Scat*) is given by

$$Scat = [N_1\ \mathrm{Var}(f_1) + N_0\ \mathrm{Var}(f_0)]/N, \tag{B.6}$$

where $N = N_1 + N_0$ is the total number of cases. *Scat* is an index of "noise" or "error," variability in forecasts that is independent of actual target event occurrences.

Minimum forecast variance. As a 1–0 indicator variable, the variance of the outcome index d is

$$\mathrm{Var}(d) = \overline{d}(1 - \overline{d}). \tag{B.7}$$

Yates (1982) showed that the following is the minimum variance in forecasts f that is possible given a particular base rate $\overline{d}$ and value of *Slope*

$$\mathit{MinVar}(f) = \mathit{Slope}^2\mathrm{Var}(d). \tag{B.8}$$

Covariance decomposition. Yates (1982; see also Yates & Curley, 1985) showed the following:

$$\overline{PS} = \mathrm{Var}(d) + \mathit{MinVar}(f) + \mathit{Scat} + \mathit{Bias}^2 - 2[\mathit{Slope}][\mathrm{Var}(d)]. \tag{B.9}$$

References

Alpert, M., & Raiffa, H. (1982). A progress report on the training of probability assessors. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 294–305). New York: Cambridge University Press.

Armstrong, J. S. (2001). Evaluating forecasting methods. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 443–472). Boston: Kluwer.

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting, 8*, 69–80.

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

Camerer, C. F., & Johnson, E. J. (1997). The process-performance paradox in expert judgment: How can experts know so much and predict so badly?. In W. M. Goldstein, & R. M. Hogarth (Eds.), *Research on judgment and decision making* (pp. 342–364). Cambridge: Cambridge University Press.

Connolly, T., & Dean, D. (1997). Decomposed versus holistic estimates of effort required for software writing tasks. *Management Science, 43*, 1029–1045.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571–582.

Enis, C. R. (1995). Expert-novice judgments and new cue sets: Process versus outcome. *Journal of Economic Psychology, 16*, 641–662.

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

Kendall, M. G. (1948). *Rank correlation methods*. London: Griffen.

Lai, K. S., & Pauly, P. (1992). Random walk or bandwagon: Some evidence from foreign exchanges in the 1980s. *Applied Economics, 24*, 693–700.

Lee, J.-W., Yates, J. F., Shinotsuka, H., Singh, R., Onglatco, M. L. U., Yen, N.-S., Gupta, M., & Bhatnagar, D. (1995). Cross-national differences in overconfidence. *Asian Journal of Psychology, 1*, 63–69.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology, 12*, 595–600.

Murphy, A. H., & Brown, B. G. (1984). A comparative evaluation of objective and subjective weather forecasts in the United States. *Journal of Forecasting, 3*, 369–393.

Mussa, M. L. (1979). Empirical regularities in the behavior of exchange rates and theories of the foreign exchange market. In K. Brunner, & A. H. Meltzer (Eds.), *Policies for employment, prices, and exchange rates, Carnegie–Rochester Conference Series on Public Policy* (vol. 11, pp. 9–57). Amsterdam: North-Holland.

Önkal, D., & Muradoglu, G. (1994). Evaluating probabilistic forecasts of stock prices in a developing stock market. *European Journal of Operational Research, 74*, 350–358.

Önkal, D., & Muradoglu, G. (1996). Effects of task format on probabilistic forecasting of stock prices. *International Journal of Forecasting, 12*, 9–24.

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

Price, P. C. (1998). Effects of a relative-frequency elicitation question on likelihood judgment accuracy: The case of external correspondence. *Organizational Behavior and Human Decision Processes, 76*, 277–297.

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

Sarantis, N., & Stewart, C. (1995). Structural, VAR and BVAR models of exchange rate determination: A comparison of their forecasting performance. *Journal of Forecasting, 14*, 201–215.

Sheskin, D. J. (2000). *Handbook of parametric and nonparametric statistical procedures* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.

Spence, M. T., & Brucks, M. (1997). The moderating effects of problem characteristics on experts' and novices' judgments. *Journal of Marketing Research, 34*, 233–247.

Staël von Holstein, C.-A. S. (1972). Probabilistic forecasting: An experiment related to the stock market. *Organizational Behavior and Human Performance, 8*, 139–158.

Windschitl, P. D. (2000). The binary additivity of subjective probability does not indicate the binary complementarity of perceived certainty. *Organizational Behavior and Human Decision Processes, 81*, 195–225.

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

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

Yaniv, I., Yates, J. F., & Smith, J. E. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin, 110*, 611–617.

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

Yates, J. F. (1990). *Judgment and decision making*. Englewood Cliffs, NJ: Prentice-Hall.

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

Yates, J. F. (1998). Conceptualizing, explaining, and improving accuracy: Process models of probability judgment. *Cognitive Studies, 5*(4), 49–64.

Yates, J. F., & Curley, S. P. (1985). Conditional distribution analyses of probabilistic forecasts. *Journal of Forecasting, 4*, 61–73.

Yates, J. F., McDaniel, L. S., & Brown, E. S. (1991). Probabilistic forecasts of stock prices and earnings: The hazards of nascent expertise. *Organizational Behavior and Human Decision Processes, 49*, 60–79.

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

Received 29 October 2001