ELSEVIER

International Journal of Forecasting 12 (1996) 41–56

international journal of forecasting

Good probabilistic forecasters: The 'consumer's' perspective

J. Frank Yates[a,*], Paul C. Price[a], Ju-Whei Lee[b], James Ramirez[a]

[a]*Judgment and Decision Laboratory, Department of Psychology, University of Michigan, 525 East University Avenue, Ann Arbor, MI 48109-1109, USA*

[b]*Department of Psychology, Chung Yuan University, Chung Li, Taiwan*

Abstract

There is an established literature describing how probabilistic forecasts, and hence forecasters, should be evaluated. The present paper takes a different and heretofore neglected perspective on evaluation. It addresses how those who receive and use probabilistic predictions—forecast 'consumers'—appraise these assessments. Results indicate that there are reliable and important differences between subjective and formal evaluation principles. Among the distinctive features of common subjective appraisal strategies are: (a) an emphasis on judgments being categorically 'correct'; (b) special attention to forecast extremeness; (c) the desire for good explanations of forecasts; and (d) the sensitivity of appraisals to how pertinent information is displayed to the evaluator. Theoretical and practical implications are discussed.

Keywords: Probablistic forecasts; Subjective forecasts; Forecast evaluation; Forecast appraisal; Subjective probability; Calibration; Scoring rules

1. Introduction

Practical decision problems often—perhaps even typically—rely on forecasts. In at least some instances, those forecasts take the form of probabilistic rather than deterministic judgments. That is, rather than making a categorical assertion that a pertinent event is going to happen, the forecaster indicates that it has some specified chance of occurring, e.g. '35%'. Moreover, even when people do not decide on the basis of probabilistic instead of deterministic forecasts, one can make a good case that they should.

An especially compelling argument for probabilistic judgments, as decision analysts have long contended, is that they allow the decision-maker to make appropriate tradeoffs with his or her own personal values (e.g. Clemen, 1991; Raiffa, 1968; Weinstein and Fineberg, 1980). This is important because the decision-maker's values can differ substantially from those of other people—*including those individuals the decision-maker might consult for forecasts*. Imagine a legal consultant who is asked to render a deterministic opinion about whether her client, the defendant, would prevail in a lawsuit. The

* Corresponding author. Tel.: (313) 747-3703; fax: (313) 763-7480; e-mail: jfyates@umich.edu or pprice@umich.edu.

SSDI 0169-2070(95)00636-2

consultant's values (e.g. for being right and wrong in particular ways) might imply that a 'Win' prediction should be reported whenever she thinks there is a 40% or better chance of the defendant winning. In contrast, the client's values for the potential lawsuit outcomes might require an assessment of at least 55% to justify taking the case to trial rather than settling out of court. If the consultant simply predicts 'Win' and 'Lose' rather than the underlying likelihood assessments, the client is deprived of information that is useful for an appropriate decision.

The quality of decisions predicated on forecasts can be no better than the quality of the forecasts themselves; inaccurate forecasts tend to yield decisions with poor outcomes. For instance, the clients of legal consultants who are poor at anticipating juries' verdicts can be expected to lose a lot of money (e.g. McGraw, 1990). This observation raises numerous questions from the perspective of people who are the 'consumers' of forecasts provided by others, their 'consultants'. At a general level, these questions include ones like the following. Suppose several consultants are available. Which one should be hired for his or her opinions? Suppose the client already has a consultant. Are the consultant's forecasts good enough? Should the consultant be retained or replaced? Furthermore, what does it mean to say that the consultant's predictions are 'good' or 'poor'?

The present paper is a discussion of questions like these, as they apply to probabilistic predictions. Over 25 years ago, in a classic article whose title inspired that of the present paper, Winkler and Murphy (1968) addressed similar issues. The big difference is that Winkler and Murphy adopted a normative stance, proposing that probabilistic forecasts be evaluated according to proper scoring rules. Among other things, such rules are designed to discourage the forecaster from reporting predictions different from his or her true beliefs. As indicated above, the perspective here is that of the forecast consumer. When people are left to their own, intuitive devices, how do they appraise the quality of probabilistic forecasts? This is an important issue for a variety of reasons, as argued presently.

Although schemes like proper scoring rules in principle have numerous advantages, they are seldom used (cf. Yates, 1994). Even under the circumstances most favorable for formal rules, we can expect intuitive appraisal methods to be far more prevalent, if for no other reason than that they are easy to apply. So, on purely scientific grounds, it is essential that we gain some understanding of an activity that is so commonly undertaken by people. What is the nature of the subjective appraisal process and why does it take on that form? And consider the practical point of view. Suppose it turns out that the principles guiding subjective appraisal strategies differ from those underlying standard analytic approaches. An immediate conclusion this would suggest is that typical consumers need help in making better appraisals, that there is a real 'need' for the kinds of tools described in the literature on formal forecast appraisal. But a more cautious and perhaps wiser response to such a discovery would be different. Discrepant intuitive and formal appraisals could serve as a hint that formal approaches themselves might be deficient and should be scrutinized more closely. And then there is the perspective of the consultant who 'sells' probabilistic forecasts. Clients will be willing to hire consultants who provide them with the kinds of forecasts they personally regard as good ones, regardless of the 'objective' quality of those predictions. Pure economic interest suggests that consultants need to know what their customers consider to be 'good'.

In the remainder of this paper we examine several specific questions about subjective appraisals of probabilistic forecasts. Our aim is to pose and elaborate the questions for public scrutiny since, curiously, they have largely escaped scholarly attention. We also describe initial empirical findings. These results suggest answers to at least some of the questions we highlight and point toward approaches that eventually should yield more definitive conclusions.

2. Study 1: Overall accuracy

A host of proposals have been made for how to characterize the overall accuracy of probabilistic forecasts. Among the ones most commonly

discussed are the quadratic and logarithmic scores (see, for example, Winkler and Murphy, 1968; Yates, 1990, 1994). By far, the most popular is a form of the quadratic score sometimes known as the 'probability score' (PS), defined as follows:

$$PS = (f - d)^2 . \quad (1)$$

In Eq. (1), $f = P'(A)$ is the probability judgment for target event A (e.g. 'the defendant will win', 'the price of security S will decrease'). The indicator variable d is called the 'outcome index', which takes on the value 1 when A occurs and 0 otherwise. Metaphorically, d can be thought of as the probability judgment of a clairvoyant. From this vantage, PS indexes forecast accuracy according to the closeness between the real forecaster's assessments (f) and those of an ideal forecaster (d). The typical accuracy exhibited by a given forecaster is normally indexed by $\overline{PS}$, sometimes called the 'Brier score' (Brier, 1950), the mean value of PS over a large and, ideally, representative sample of cases. The first question we ask about subjective appraisals is how they compare with standard, formal ones, in particular, the probability score. Answering this question was one of the major aims of Study 1.

Thirty-six undergraduates served as paid subjects in the study. Two tasks were prepared, with approximately half the subjects randomly assigned to each task. In the 'Meteorologist Task', the subject assumed the role of a citizen asked by the National Weather Service (NWS) to make a recommendation of which of two meteorologists, 'Blue' or 'Green', the subject would prefer to be hired at the local NWS unit. To support such a recommendation, the subject would be provided with each meteorologist's probabilistic forecast of rain for each of 48 consecutive days, along with the actual outcome, rain or no rain. Each forecast–outcome pair was presented on a separate index card in a format equivalent to the following example:

Day 5
Judged probability of rain: 30%
Actual outcome: No rain

The subject was presented with all 48 cards for Forecaster Blue and another 48 cards for Forecaster Green. The subject was encouraged to manipulate the cards on a table and to summarize them any way he or she liked before expressing a preference between the forecasters. After stating a preference, the subject was asked to write out an explanation for his or her choice.

The 'Stock Broker Task' was the same as the Meteorologist Task except for the cover story. In the Stock Broker Task, the competing forecasters were stock brokers. They made predictions about whether each of 48 stocks would increase in value over the next financial quarter. The subject was asked to indicate which forecaster was preferable as a personal financial advisor and to explain that preference.

The two different data sets, i.e. collections of 48 forecast–outcome pairs, were constructed in a special way. For reasons described below, one set is labelled 'Data Set C', the other 'Data Set D'. Henceforth, we will describe the corresponding forecasters as Forecasters C and D, respectively. Since subjects' responses were the same regardless of whether they performed the Meteorologist Task or the Stock Broker Task, all responses have been collapsed across cover stories. Data Sets C and D differed systematically with respect to two specific accuracy dimensions, as discussed in a subsequent section of this paper. For present purposes, however, the important feature of the data sets is that the overall accuracy of the forecasts they contained was the same, as indexed by $\overline{PS}$, approximately 0.19 in each instance.[1] Accordingly, if subjective appraisals of overall forecast accuracy correspond to the concept of accuracy embodied in the probability score, then on average subjects

[1] Constant judges provide useful reference points for interpreting probability scores. A constant judge who reports the same judgment $f = c$ for every case that comes along would achieve a mean probability score of $\overline{PS} = \bar{d}(1 - \bar{d}) + (c - \bar{d})^2$. Thus, the best possible constant judge is one who always indicates the base rate $\bar{d}$ as his or her forecast. In this study, since $\bar{d} = 0.42$ approximately, such a base rate judge would earn a score of $\overline{PS} = 0.2436$. Another special constant judge is the 'know-nothing' uniform judge who consistently says that the target event is just as likely to occur as to not occur, achieving $\overline{PS} = 0.25$ with no effort at all.

should have been indifferent between Forecasters C and D. In fact, they were far from indifferent. The great majority, 28 (78%) of the 36, chose Forecaster D. Given a null hypothesis of no consistent preference between the two forecasters (i.e. for each subject there was a 0.5 probability of preferring either forecaster), this result is extremely unlikely ($p = 0.001$ via a binomial test).

3. Study 2: Overall accuracy again

The results of Study 1 indicate that subjective appraisals of overall accuracy are derived according to principles different from those implicit in the probability score. Study 2 was an open-ended exploration intended to seek hypotheses for what such principles might be. The subjects were 44 undergraduate, graduate, and professional students in a course on decision processes. At the time, the students knew what probabilistic forecasts were; they recently had completed an intensive forecasting exercise themselves, and the actual outcomes of the focal events had just been determined. However, the students were still ignorant about formal techniques for evaluating the accuracy of such assessments. The pedagogical aim of the project constituting this study was to provide an orientation for learning about these methods.

In the 'Intuitive Analysis Strategy' project, each subject was asked to do the following:

> Imagine that you are a vice president of a fairly large corporation. You have been charged with selecting a law firm that will be put on a retainer for handling your company's litigation (i.e. various kinds of lawsuits). Numerous considerations bear on your decision. But one of them is how good a given law firm is at anticipating how lawsuits would be decided... . Thus, you ask each firm to perform a task that is analogous to last week's... forecasting exercise. The firm was allowed to gather as much information as it desired about each of 100 pending lawsuits, which were similar to ones that occur in your business. In each case, the firm was asked to predict who would win the suit (they all went to trial), the plaintiff or the defendant. The firm then stated a 50%–100% probability judgment that the predicted winner really would win.

The subject was then instructed: "Based on your own personal intuitions about what seems appropriate..., outline the strategy you would use in evaluating the accuracy of each firm that took part in the exercise". Furthermore, the subject was asked to indicate exactly how he or she would determine whether a firm's judgments were "good or poor, and in what particular ways". Subjects had a week to develop their proposals.

Several recurring themes emerged from the subjects' responses. We summarize and comment on those that are especially pertinent to the present issues (frequencies of mentions are denoted by N in parentheses).

Process and explanation ($N = 17$): One of the most popular considerations mentioned by subjects was the process by which forecasters arrived at their judgments. As Subject 28 put it, "In order to assess the accuracy of predictions by the different law firms, one would need a documentation of the strategies or methods...employed by these firms in order to arrive at a probability judgment." Getting more specific, Subject 27 said, "I would want to take into account what information and how much information they used to make their judgments." Many of the subjects who maintained that process was important also reported that they would seek process information directly from the forecasters themselves, from their explanations. Thus, Subject 14 said she would "be sure to ask each firm to describe its forecasting strategy in addition to its trial predictions". Subject 44 indicated that she would "request from the firms a detailed explanation of why certain probabilities were assigned".

One reason these remarks are significant is that they speak to what consumers regard as adequate justification for relying on a consultant. From the perspective implicit in the empirical

tradition, it should be irrelevant how a consultant arrives at his or her assessments, only that those judgments are reliably good in a statistical sense (e.g. based on a large sample of 100 cases). But that is apparently not good enough for many consumers. As some of the present subjects said explicitly, they did not want to be misled by a run of good luck on the forecaster's part; they wanted proof that a forecaster adheres to a forecasting method that can be publicly justified. This disposition is consistent with the experience of developers of systems for performing various judgment and decision tasks. For instance, early Bayesian inference systems simply cranked out posterior probabilities for users and that was that. Such systems were never very popular. Designers of later expert systems, which are now fairly common in areas like medicine, argue that part of the reason for user 'coolness' to the earlier efforts was that they did not provide natural language explanations. Hence the current practice of offering such justifications (e.g. Buchanan and Shortliffe, 1984).

The literature suggests that the insistence by some consumers that forecasters explain their own forecasts is not without risks. Research indicates that people sometimes have limited insight into how they arrive at their judgments and decisions, and that such 'self-insight' is especially poor for those who are expert at a given task (e.g. Slovic et al., 1972). In the research literature, people are said to be 'accountable' when they expect that they might be called upon to justify their judgments or decisions. Studies have shown that accountability sometimes leads to reliably improved judgments (e.g. Tetlock, 1985), but at other times consistently worse performance (e.g. Siegel-Jacobs and Yates, 1995; Simonson, 1989). Implicitly, such results suggest that what is (thought to be) convincing to other people is not necessarily the same as what is 'right'. And then there is the matter of the consumer's ability to understand the rationales a forecaster might offer and also the forecaster's articulateness. There are likely to be many instances in which consumers are unable to evaluate explanations competently. Similarly, it is easy to imagine forecasters who are highly accurate but inarticulate, and vice versa (see also Teigen, 1990).

50% aversion ($N = 8$): Another theme that emerged in several subjects' intuitive analytic schemes was a disdain for 50% forecasts. Subject 11 put it this way: "I personally have no use for a 50% forecast, I can say that I don't know all by myself." Subject 27 offered another explanation for his negative feelings about such assessments: "A firm who most often made low probability judgments, at or around 50% most often, probably has not gathered enough information to be sure of themselves, and is not the firm for me." In general, subjects who eschewed 50% forecasts did so for one or more of several reasons. They took such judgments as indications that the forecasters were either generally incompetent, ignorant of the facts in a given case, or lazy, unwilling to expend the effort required to gather information that would justify greater confidence.

These arguments are not without merit. But there are at least two considerations they might neglect. First of all, although 50% judgments often do indeed reflect deficient forecaster performance, this is not always the case. Sometimes the forecaster has gathered and properly evaluated all the available information, but that information is simply contradictory. Such ambiguity is one of the weaknesses in probabilistic expression that motivated the emergence of alternatives such as belief functions (e.g. Shafer, 1976). To date, however, to the best of our knowledge, there are no well-developed techniques for analyzing the quality of belief functions as forecasts of real-world occurrences. Given the popularity of belief function approaches in artificial intelligence, such analytic methods should be a priority for future research. Consumer aversion for 50% forecasts might also neglect the need to verify the legitimacy of more extreme assessments. If an unscrupulous forecaster recognizes that a consumer holds 50% judgments in contempt, the forecaster might assign more extreme judgments arbitrarily, hoping and perhaps anticipating that the consumer will be lax about documenting that such judgments are justified.

Categorically 'correct' forecasts ($N = 13$): Numerous subjects placed special significance on forecasters correctly predicting events in a categorical sense. In the present exercise, forecasters were described as having provided their predictions according to a two-stage process. First, the forecaster indicated whether he or she expected the plaintiff or the defendant to win a given lawsuit. The forecaster then stated a 50%–100% probability that that expectation would be borne out in the trial. Thus, a 'correct' forecast was the selection of the actual winner of the case. More generally, within this perspective, a correct forecast would be the assignment of a probability of more than 50% to the event that actually happens when only two alternatives are specified. Subject 43 expressed his disposition about correct forecasts this way: "First and foremost, the firm should predict the correct winner most of the time. The deterministic portion of the judgment, i.e. who will win, is the best indicator of judgment accuracy, so a poor percentage of correct predicitons would not bode well." And Subject 36 submitted that "The number of correct and incorrect responses is valid and obvious proof (attesting) to a firm having good judgment."

Once again, the appeal of categorical correctness seems not unreasonable, on its face, at least. On closer scrutiny, however, it might implicate some problematic issues. If forecasts are expressed probabilistically, there is no inherent need to attach any special significance to judgments in any particular category (e.g. greater or less than 50%). As argued at the outset of this paper, in the discussion about the usefulness of probabilistic forecasts, the importance of judgments in certain ranges should depend on tradeoffs between probabilities and the consumer's values for the outcomes of various actions predicated on the forecasts. Thus, special attention to 'correct' and 'incorrect' judgments—even their mere designation—might suggest that the forecast consumer does not fully appreciate or acknowledge the advantages of probabilistic as opposed to deterministic forecasts. Such a disposition might reflect the greater familiarity of deterministic forecasts. It also seems plausible that it might be peculiar to situations in which probabilistic forecasts are elicited in a two-stage scheme in which the first stage requires an explicit deterministic prediction (e.g. which party will win a lawsuit). We return to this latter possibility below.

Probability-proportional scoring ($N = 12$): Consider the following scoring scheme proposed by Subject 9:

> For each correct response (i.e. predicting a winning suit would win or that a losing suit would lose) the probability percentages would be added. From this score, the probability percentages of each incorrect response would be subtracted. This would give a score from −100 to +100 for comparison.

Or take the equivalent proposal as articulated by Subject 21:

> If the firm was correct in its prediction and gave a 100%, the firm would score a 5, 90% a 4, 80% a 3, 70% a 2, 60% a 1, 50% a 0. If the firm is wrong, it receives from 0 to −5 depending on the probability judgment. If the firm gave a 100% wrong prediction it receives a -5, 90% a -4, 80% a -3, 70% a -2, 60% a -1 and 50% a 0.

This kind of scoring routine was offered by a number of subjects. The essence of the method is that a score is assigned to a forecast in proportion to the probability attached to the event that eventually actually occurs.

As with all the considerations mentioned by the subjects, there is a certain sensibleness to the proportionality scheme: after all, why should forecasters not be rewarded in direct relation to the probabilities they give to what ultimately happens? Indeed, the proposed approach is equivalent to one that has been discussed on several occasions in the literature (e.g. Winkler, 1969), namely the 'linear scoring rule' (Lr), which can be defined as follows:

$$Lr = fd + (1 - f)(1 - d)\,. \tag{2}$$

Recall that the notation is such that $f = P'(A)$ is

the probabilistic forecast for target event A, $1-f=P'(A^c)$ is the corresponding forecast for the complementary event A^c, and outcome index d takes on the value 1 when A occurs and 0 when A^c is observed. This rule yields a score that is identical to the probability given to the event that really happens.

A significant feature of the linear score is that it is not 'proper', in the sense discussed by Winkler and Murphy (1968). In fact, it can be shown (cf. Yates, 1990, pp. 72, 239) that it is in the interests of a forecaster compensated according to the linear score to misrepresent his or her true beliefs in a particular way. Specifically, if the forecaster believes that an event is more likely to occur than not, then an extreme judgment of 100% should be reported for that event. Although this property of the linear score is not immediately obvious (to wit, our subjects' recommendation of it), research suggests that individuals rewarded according to the linear score eventually come to recognize its nonproperness, at least implicitly in their actions (Jensen and Peterson, 1973).

Recall that subjects in Study 1 strongly preferred Forecaster D to Forecaster C. Among other things (as discussed below), this preference is consistent with evaluation according to a linear scoring rule. It turns out that the mean linear score for Forecaster D was 0.71 while that for Forecaster C was 0.64.

4. Study 1 revisited: Accuracy dimensions

Study 2 suggests that at least some of the preferences in Study 1 for Forecaster C vs. Forecaster D might have been based on overall accuracy assessments guided by principles describable by schemes like the linear scoring rule. But the responses of subjects in Study 2 also suggest another possibility. Some of the Study 1 preferences might have been driven by evaluations of Forecasters C and D on more specific 'accuracy dimensions', at the level of constructs such as an aversion for 50% judgments. Study 1 had been designed to test this possibility as well as the extent to which forecaster appraisals agreed with the prescriptions of $\overline{PS}$.

4.1. Standard dimensions

The forecast–outcome data sets for Forecasters C and D were designed such that they differed in a particular way with respect to two 'standard' accuracy dimensions while being equivalent in overall accuracy according to $\overline{PS}$. Those dimensions are most easily perceived and understood with the aid of calibration graphs. Fig. 1 shows the graphs for the assessments provided by Forecasters C and D.

The abscissa in each graph shows the probabilistic forecasts, in tenths. The ordinate describes the corresponding proportions of times the target event occurred when a particular judgment or forecast was made. Thus, consider the point above 0.9 in Forecaster C's graph. The number 4 next to that point indicates that there were four instances in which Forecaster C said that there was a 90% chance that the target event (i.e. 'rain' or 'stock price increases') would occur. (In some calibration graphs, like those shown here, in order to convey the proper weighting of particular judgments, the areas of points are drawn proportional to the frequencies with which the pertinent forecasts were offered.) The ordinate of that point is at 0.75, indicating that the target event actually occurred in three out of those four instances. 'Calibration' refers to the extent to which probability judgments match the corresponding proportions of event occurrences. In a calibration graph, good calibration is indicated by the closeness of large points to the 1:1 diagonal. It is encoded numerically by the 'calibration index' (see Yates, 1990, ch. 3; 1994, for exact expressions for the various indexes discussed here). As suggested by Fig. 1, the calibration of Forecaster C was superior to that of Forecaster D; hence the label 'C'.

Forecasts exhibit good 'discrimination' to the degree that the forecasts reported for instances when a target event is ultimately going to occur are different from those reported when it is not. It is important to recognize that, while the calibration concept is intimately wedded to the

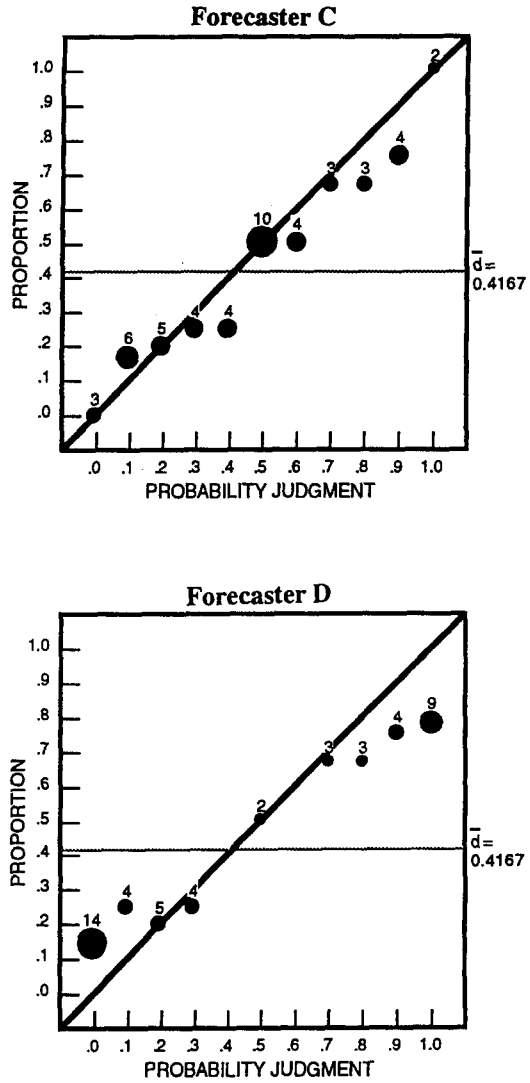

Fig. 1. Calibration graphs of the forecast–outcome data used in the present studies.

numerical character of the forecaster's predictions, the numbers per se are immaterial to discrimination. All that matters to good discrimination is that the forecaster is disposed to report *different* probabilities for target-positive and target-negative cases. In a calibration graph, good discrimination is indicated when large points (representing heavily used forecast categories) have vertical coordinates that are maximally different from the horizontal line identified with the 'base rate', $\bar{d}$. The base rate for the target event is simply the overall proportion of times the target event occurs, e.g. about 41.7% for 'rain' and 'stock price increases' in Study 1. As Fig. 1 indicates, and as confirmed by the calculated 'discrimination indexes' (see Murphy, 1973; Yaniv et al., 1991; Yates, 1990, ch. 3; 1994), Forecaster D's forecasts were more discriminative than Forecaster C's; thus the 'D' characterization.

An alternative 'standard' approach to decomposing overall probabilistic forecast accuracy into dimensions entails 'covariance graphs', as illustrated in Fig. 2 for Forecasters C and D (Yates, 1982, 1990; Yates and Curley, 1985; see also Murphy and Winkler, 1992). A covariance graph consists of two histograms of forecasts for the target event. The one on the left, above the outcome index value $d = 0$, describes the forecasts for those occasions when the target event failed to occur; the one on the right ($d = 1$) displays forecasts for instances when it did occur. The ordinate of the graph describes the various probabilistic forecasts that might have been rendered. The abscissa describes the alternative values of the outcome index d as well as the possible values of the base rate $\bar{d}$, scaled between the extremes of 0 and 1. Ideally, all the markers in the histogram on the left should be at $f = 0$, while all those in the histogram on the right should be located at $f = 1$. 'Covariance decomposition analysis' focuses on three particular ways judgments can fall short of this ideal.

The first dimension isolated in covariance decomposition analysis is 'bias', the difference between the mean forecast overall and the base rate: bias $= \bar{f} - \bar{d}$. In a covariance graph, the bias is easily discerned as the vertical distance between the 1:1 diagonal and the intersection of the horizontal $\bar{f}$ line and the vertical base rate, $\bar{d}$ line. The bias is positive, indicating overprediction of the target event, when the intersection is above the diagonal; it is negative when the intersection occurs below the diagonal. As Fig. 2 shows, the biases of both Forecasters C and D were positive, but that of Forecaster C was larger.

The second dimension distinguished in covar-

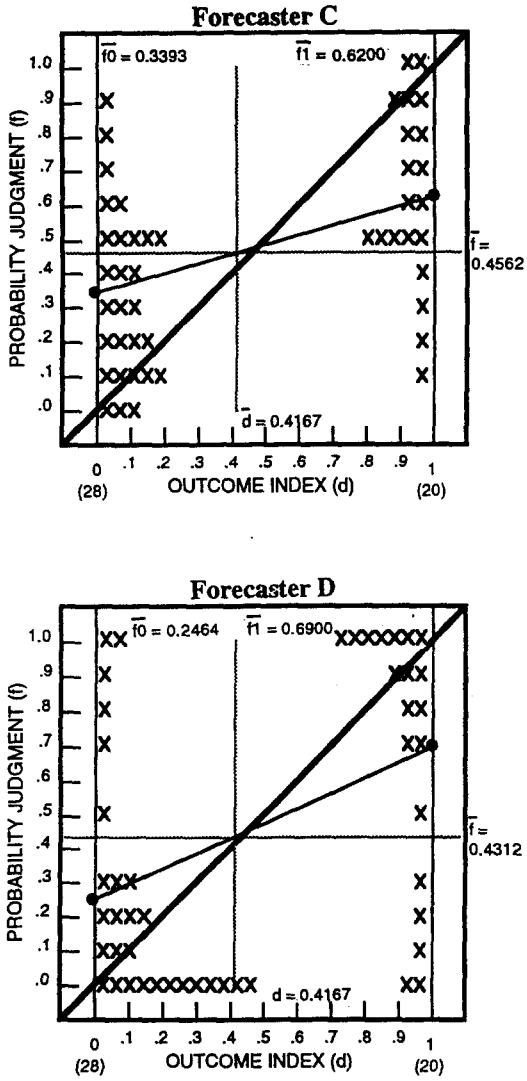

Fig. 2. Covariance graphs of the forecast–outcome data used in the present studies.

iance decomposition analysis is slope, the difference between the mean forecast ($\bar{f}_1$) for the target event for the occasions when it actually happened ($d = 1$) and the corresponding mean ($\bar{f}_0$) for the occasions when it failed to occur ($d = 0$): slope $= \bar{f}_1 - \bar{f}_0$. Ideally, this statistic should be 1. Literally, this measure is the slope of the regression line of forecasts on the outcome index, the line that passes through the points in the two histograms in the given covariance graph. Fig. 2 makes it immediately apparent that the slope of Forecaster D was better than that of Forecaster C.

The final dimension covariance decomposition analysis examines is scatter, the amount of variability contained in the forecaster's assessments that is unrelated to the occurrence of the target event, a kind of 'noise' or random variation. Graphically, scatter is represented by the variability within the two histograms contained in the pertinent covariance graph. It is indexed by a scatter statistic that is a weighted average of the variances of the forecasts in the two histograms. As is probably apparent from inspection of Fig. 2, the scatter in Forecaster D's predictions was greater (i.e. worse) than that in Forecaster C's.

4.2. Preference implications

The primary dimensional focus in Study 1 was on whether forecast consumers had stronger preferences for good calibration or for good discrimination in situations where, from an 'objective' point of view, these two dimensions offset each other. Murphy (1973) has shown (see also Yates, 1982, 1990, 1994) that calibration and discrimination indexes compensate for each other in a decomposition of $\overline{PS}$ similar to the partitioning of a sum of squares in the analysis of variance. This Murphy decomposition is what permitted the construction of data sets such that Forecaster C's weak discrimination relative to that of Forecaster D could be offset almost exactly by stronger calibration. The subjects' clear preference for Forecaster D (78% vs. 22%) is consistent with a conclusion that forecast consumers value good discrimination more than good calibration. A good case can be made (e.g. Yates, 1994) that this is as it ought to be. Depending on its form (e.g. consistent over- or underprediction), poor calibration sometimes can be improved by relatively simple means, including mathematical translation of forecasts after the fact. But good discrimination can be achieved only when the forecaster has access to good, diagnostic information and knows how to interpret it.

Recall that Forecaster D's forecasts were su-

perior to those of Forecaster C in other respects besides discrimination. In particular, those predictions were less biased and had better slope, although they were more scattered. Thus, the preference for Forecaster D is also consistent with forecast consumer emphases on unbiasedness or slope. Besides stating their preferences between Forecasters C and D, subjects in Study 1 were asked to perform a second task to provide insight into which accuracy dimensions really did affect their forecaster choices.

4.3. Dimensional accuracy recognition

After subjects chose between Forecasters C and D and wrote explanations for their choices, the forecast–outcome cards were removed. The subjects were then provided with an Accuracy Dimensions Questionnaire. The questionnaire first provided explanations (without graphs) of the standard dimensions of bias, calibration, discrimination, slope, and scatter. It then asked the subject to indicate whether Forecaster C or Forecaster D had performed better on each of those dimensions. Subjects also rated their confidence that the choices they made were the correct ones, on a nine-point numerical rating scale (ranging from 1 to 9), anchored with the verbal labels 'Guessing' and 'Extremely Confident'. The reasoning behind this memory-based task was that subjects should be able to identify the forecaster who was actually better on a given dimension only if they had attended to (and perhaps relied on) that dimension during the prior preference task.

For each of the five standard accuracy dimensions we performed a t-test on subjects' ratings of their confidence that they chose the better of the two forecasters. Before doing so, however, we recoded subjects' confidence ratings in the following way. First, we subtracted 1 from each confidence rating so that the ratings of subjects who felt that they were guessing were now 0, and the ratings of subjects who were extremely confident were 8. Then we multiplied the ratings by -1 whenever the worse forecaster on the accuracy dimension in question was chosen as the better one. Thus, a positive confidence rating indicated a correct choice and a negative rating implied an incorrect choice.

Subjects' mean confidence ratings for the five standard accuracy dimensions, along with the p-values resulting from the t-tests, are presented in Table 1. Note that subjects had a tendency to believe that Forecaster D outperformed Forecaster C on every accuracy dimension. (The actual dimensional indexes are included in Table 1 also.) Furthermore, this tendency was statistically significant (i.e. the mean was significantly different from zero at the 0.05 level) for the dimensions of calibration, discrimination, and slope. This is surprising in the case of calibration, since the forecast–outcome pairs were generated expressly so that the calibration of Forecaster D was worse than that of Forecaster C.

One plausible explanation for these results is that subjects based their preferences on one particular accuracy dimension, perhaps discrimination or slope, and then simply responded to the Accuracy Dimensions Questionnaire in a manner that was consistent with their stated preferences. For example, a subject may have

Table 1
Values of standard accuracy dimenson indexes for Forecaster D (good discrimination) and Forecaster C (good calibration), along with subjects' mean confidence ratings that they correctly identified the better forecaster on each accuracy dimension in Study 1, and the p-values associated with the mean confidence ratings (H_0: mean confidence = 0)

Dimension	Fcstr. D	Fcstr. C	Conf.	p-value
$\overline{PS}$[a]	0.1869	0.1896		
Calibration index[a]	0.0264	0.0073*	−2.56	0.0069
Discrimination index[b]	0.0826*	0.0608	3.42	0.0001
Bias[c]	0.0104*	0.0375	0.28	0.7579
Slope[b]	0.4507*	0.2757	2.56	0.0003
Scatter[a]	0.1134	0.0607*	−1.17	0.1907

[a] Smaller values are better.
[b] Larger values are better.
[c] Closer to zero is better.
Note: Asterisks indicate the better judge on each accuracy dimension. Confidence ratings were positive when the subject correctly chose the better forecaster on each accuracy dimension and negative when the subject chose incorrectly. All p-values are for two-tailed t-tests, df = 35.

attended to nothing except discrimination, noticed that Forecaster D had better discrimination, and then stated a preference for Forecaster D. On the subsequent Accuracy Dimensions Questionnaire, the subject may have had no idea which forecaster had better calibration, but inferred that it was Forecaster D because of his or her overall preference for Forecaster D. Responses on the Accuracy Dimensions Questionnaire may, in effect, reflect a halo effect (Cooper, 1981).

4.4. Rationales and intuitive dimensions

Recall that after subjects stated their preferences between Forecasters C and D, and before they responded to the Accuracy Dimensions Questionnaire, they provided rationales for their preferences. These rationales were subjected to a content analysis by two coders blind to the aims of the study. The coders examined the rationales for indications that the subjects referred to each of the five standard accuracy dimensions discussed above. They also coded the presence or absence of references to five additional 'intuitive' dimensions: 'number of correct judgments', 'judgment extremeness', 'deviation when incorrect', 'judgment utility', and 'number of better judgments'. The existence of these dimensions was suggested by an initial reading of the rationales by the investigators. Working independently after training, the coders achieved an agreement of 84%. They then re-read the explanations together to eliminate their disagreements. Table 2 presents a summary of the results, including quotations illustrating each of the dimensions, standard and intuitive, which are discussed more fully below.

Perhaps the most striking finding revealed in Table 2 is the rareness of references to constructs equivalent to the standard accuracy dimensions. Only 16% of the total number of references in the rationales concerned the standard accuracy dimensions; the rest were references to intuitive considerations. It is especially noteworthy that the subjects never mentioned anything that could be construed as a reference to either calibration or discrimination. The dimensions isolated in covariance decomposition analysis fared only slightly better. Although the standard dimensions are compelling on theoretical grounds, they apparently are not foremost in the minds of

Table 2
The number and percentage of subjects (out of 36) who mentioned each accuracy dimension – along with an example for each dimension – in explaining their preferences between the forecasters in Study 1

Dimenson	Number	Pct.	Example
Number of correct forecasts	23	63.9	"... Mr. Green ... when he said he was more than 50% sure he was correct more often than Mr. Blue."
Fcst. extremeness	20	55.6	"... Green's probability estimates tended towards the extremes while that of blue tended toward the medians."
Deviation when incorrect	10	27.8	"The blue forecaster made less drastic mistakes."
Slope	6	16.7	"If there was a probability of rain, he/she judged a high probability (80%–100%) or if ... it didn't rain ... Green judged a low probability (0%–20%)."
Judgment utility	3	8.3	"When an assertion like 0% probability is made, it conveys certain implications for action that may be catastrophic if the event ... actually occurs."
Scatter	3	8.3	"Blue ... seemed to be all over w/their predictions."
Number of better judgments	2	5.6	"So if the actual weather ... was rain, I looked for the higher percentage. The higher percentage was the winner. If the actual weather was no rain, I looked for the lower percentage."
Bias	2	5.6	"Green did overestimate the probability of rain, but ... less than Blue."
Calibration	0	0.0	
Discrimination	0	0.0	

forecast consumers, at least in a context that is faithful to real-world forecasting situations in which forecasts and outcomes are presented one case at a time.

Observe that the most frequently mentioned intuitive dimension, number of correct judgments, was one of those cited by subjects in Study 2, who were asked to describe in the abstract how they would approach the task of appraising probabilistic forecasting accuracy. It is significant that this dimension emerged so prominently even here, where (in contrast to Study 2) the forecasting task did not entail a two-staged procedure in which the forecaster first reported a categorical prediction (e.g. 'rain' vs. 'no rain'). Instead, Forecasters C and D had indicated probabilities for a single designated target event (e.g. 'rain'). It is worth noting that, using either strict or lenient definitions of 'correct', Forecaster D's predictions were better than those of Forecaster C. A correct prediction in the strict sense is an indication of 100% certainty in an event that ultimately occurs; according to the lenient sense, a forecast above 50% for an event that eventually happens is considered correct.

The second most commonly cited dimension was 'judgment extremeness'. Again, this dimension is essentially the same as the one emphasized by subjects in Study 2. There it was described as an aversion to 50% forecasts. Forecaster D was better than Forecaster C with respect to this dimension, too. For instance, the mean absolute deviations from 50% of the forecasts by Forecasters D and C were 38% and 23%, respectively.

The remaining, less frequently mentioned intuitive dimensions were not brought up by subjects in Study 2. 'Deviation when incorrect' refers to cases when the actual event that occurred was not the 'correct' one in the senses described above. A large deviation or 'drastic error' took place when the forecaster stated a very high probability for the 'wrong' event. Interestingly, a reliance on deviation when incorrect would have supported a choice of Forecaster C over Forecaster D. The mean forecast–outcome index deviation was 73.3% when Forecaster C was incorrect, but this statistic was 88.2% for Forecaster D. 'Forecaster utility' refers to potential practical implications of forecasts. And the dimension 'number of better forecasts' pertains to pairwise competitions such as that staged between Forecasters C and D, where both competitors make predictions for the very same cases. Reliance on this dimension requires that the consumer keep a tally of cases for which one forecaster's predictions more faithfully anticipate the actual outcome than do the other's. The forecaster with the higher tally is considered the 'winner' and hence the better judge.

5. Study 3: Extremeness preferences

Numerous subjects in both Study 1 and Study 2 indicated that an important consideration to them in evaluating probabilistic forecasters is the extent to which those forecasters assign either middling judgments around 50% or, conversely, extreme judgments near 100%. Taken as a whole, the indications are rather convincing that what we can call the 'extremeness' dimension is a significant one for many judgment consumers. Nevertheless, it would be helpful to have even more evidence, evidence that is not clouded by such factors as the correlation in Study 1 between extremeness and other accuracy dimensions in the forecast–outcome pairs of Forecasters C and D.

To achieve a 'cleaner', unconfounded test of the role of the extremeness dimension in subjective evaluations of forecasters, we employed the same basic approach as in Study 1, but with a few critical changes. Thus, a new sample of 28 undergraduate subjects was presented with the 48 forecasts for Forecasters C and D and then asked to indicate which they would prefer and why. To simplify matters, only the stock broker cover story was used. Also, subjects were not only asked to indicate which forecaster they preferred, but to also rate the strength of their preferences on a 10-point rating scale ranging from 0 to 9. Because it was deemed important to allow subjects to express indifference in this study, the rating scale was anchored with the

labels 'No Preference' and 'Very Strong Preference'.

The most crucial difference between this study and Study 1 is that here the subjects were not told what the actual outcomes were, i.e. whether the focal stock in a given case did or did not eventually increase in price. The main reason for this feature of the design was that it would eliminate correlations between extremeness and the other accuracy dimensions discussed earlier. Although the withholding of outcome information might seem bizarre, it actually has considerable ecological validity. There are many real-world circumstances in which consumers learn about the outcomes of the events forecasters are asked to predict only after long periods of time, if they ever learn about the outcomes at all. For instance, medical patients individually see their physicians only occasionally, and they almost never get definitive evidence of what actually ails them. Thus, they do not have the means to evaluate the quality of those physicians' diagnostic skills. Indeed, we are unaware of any field—including medicine, clinical psychology, financial services, law, and, to a lesser extent, meteorology—in which advisors spontaneously and routinely offer to clients the records that are required to assess their accuracy. In effect, all that judgment consumers often get are their consultants' judgments themselves.

Fifteen subjects (53.6%) preferred the forecaster with more extreme predictions, Forecaster D; 12 (42.8%) picked the more moderate Forecaster C; and one subject (3.6%) was indifferent. These preferences are not reliably different from what one would expect if there was a 0.5 probability that a randomly sampled subject preferred an extreme rather than a moderate forecaster ($p = 0.84$ per binomial test).

We multiplied by -1 the ratings of subjects who chose the moderate forecaster. Thus, positive transformed ratings indicate a preference for more extreme forecasts, and negative ratings a preference for moderate predictions. The mean rating (0.71) was not significantly different from 0, $t(27) = 0.55$, $p = 0.59$, consistent with the frequency analysis. A visual display of the distribution of subjects' preference ratings, as shown in Fig. 3, is more informative, though. Observe that the distribution is distinctly bimodal. Thus, it is inappropriate to conclude that the typical, individual subject was indifferent between the extreme and moderate forecasters. On the contrary, the individual preferences were, with very few exceptions, quite strong. It is just that subjects who liked the moderate forecaster were just as common as those who disliked that forecaster.

An analysis of subjects' rationales for their choices (as encoded by two independent, trained coders) suggested why the preference patterns took the form they did. A number of themes were revealed in subjects' explanations. But two of them predominated and seem especially pertinent. The first can be called 'certainty', and is illustrated by the following quotation: "I chose blue because she seemed confident on the ones that would increase in value, and the ones that wouldn't." It is noteworthy that all 10 of the subjects who cited certainty preferred the more extreme forecaster, Forecaster D. The second major theme can be labeled 'realism', and is exemplified by this rationale: "...green...was much less likely to make definite 'yes-this-one-is-

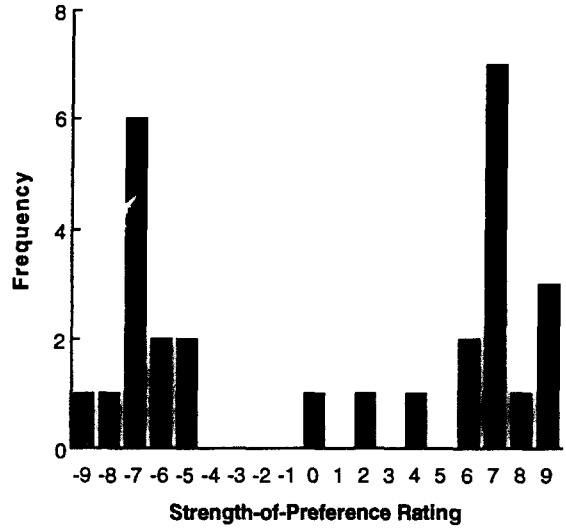

Fig. 3. The distribution of subjects' strength of preference ratings in Study 3. Positive ratings indicate a preference for the extreme forecaster (Forecaster D), and negative ratings indicate a preference for the moderate forecaster (Forecaster C). A rating of zero denotes indifference.

a-winner' predictions... . This seems much more realistic... ." All but one of the seven subjects who mentioned realism picked the more moderate forecaster, Forecaster C.

Results like these suggest that extremeness is indeed a consideration that many forecast consumers take into account when they evaluate forecasters. But the interpretation they give to such extremeness is subject to marked individual differences. For some consumers, extremeness is taken as an indicator of a positive quality, confidence in one's convictions. There seems to be an implicit assumption by such consumers that a confident forecaster would not express such certainty unless it were justified. Other consumers are more skeptical, suspecting that extreme forecasters might know so little about the inherent uncertainty in the situation that they fail to realize that they are being reckless. These dual and opposing interpretations of the same 'cue' about forecaster competence seem closely related to Teigen's (1990) findings that people use the very same signs to draw the conclusion that an informant is an expert as that he or she is a fraud.

6. Study 4: Graphical displays

The previous studies have all applied to situations in which the forecast consumer is essentially a 'layperson' with respect to analytical methods for evaluating forecast accuracy. But suppose the consumer has been instructed about the meanings of the 'standard' accuracy dimensions and even provided with pertinent calibration or covariance graphs. Of special interest is whether such informed consumers' preferences for forecasters are affected by whether forecast–outcome data are displayed in the form of calibration as opposed to covariance graphs. Study 4 was designed to address this issue.

The 29 subjects who served in Study 4 were undergraduate, graduate, and professional students enrolled in a decision processes course. They participated in the study as part of an in-class exercise. Prior to the exercise, the students had been taught how to read and interpret calibration and covariance graphs. However, they had not been exposed to arguments about tradeoffs among various accuracy dimensions, such as calibration, discrimination, and slope, i.e. whether and why a forecast consumer should prefer strength on one dimension rather than another.

As in Studies 1 and 3, the basic task for the subject was to express a preference between Forecasters C and D. The major difference between this study and Study 1 was that here subjects were not presented with forecast–outcome pairs one at a time. Instead, in each of two separate tasks, they saw the data sets as summarized in either calibration graphs or covariance graphs. Thus, in one version of what we could call the 'Calibration Graph Task', a subject might be told that Forecasters C and D were the NWS meteorologists described in Study 1. The subject was presented with the calibration graphs (as in Fig. 1) for the two different forecasters (with no calibration and discrimination indexes provided), side-by-side. The subject was then allowed 15 min to examine the graphs, indicate which forecaster was preferred, and to write a brief explanation of the preference. At the next class meeting, the same subject was asked to perform the 'Covariance Graph Task'. This time, the stock broker cover story of Study 1 was used. Unbeknown to the subject, the data sets were the very same ones he or she had compared in the previous session. Only this time, the data were displayed (as in Fig. 2) in covariance graphs (with no indication of bias, slope, and scatter measures). Care was taken to counterbalance the left–right positioning of graphs and the order of the tasks.

Subjects' preferences for Forecaster D, the forecaster with better discrimination, can be summarized as follows:

Calibration Graph Display: 57.7% (15 out of 26 subjects).

Covariance Graph Display: 96.0% (24 out of 25 subjects).

Thus, we see that the form of graphical presentation had a dramatic effect on subjects' preferences between Forecasters C and D. When the

subjects examined calibration graphs, the preference ratio was not significantly different from 0.5 ($p = 0.56$, binomial test). But when the subjects were presented with the corresponding covariance graphs, that ratio was markedly different from 0.5 ($p = 0.000001$, binomial test). Following McNemar (1969), we found that the number of subjects who preferred Forecaster C when given calibration graphs but changed to a preference for Forecaster D when shown covariance graphs (8) significantly exceeded the number (0) who exhibited the opposite change in preference ($p = 0.008$, binomial test).

We thus see that the form in which forecast–outcome correspondence is displayed does indeed greatly influence forecast consumers' preferences between forecasters. But why did we observe the specific effects that emerged? The explanation probably rests at least partly on the qualities particular graphs make salient. Calibration graphs, for instance, bring immediate attention to calibration, even for naive observers. This effect plausibly contributes to the inordinate emphasis on calibration rather than discrimination in the research literature on subjective judgment; calibration graphs have been far more common than other forecast displays. In contrast, covariance graphs make it especially easy for an observer to recognize distinctions in bias and in slope, the latter of which is often strongly related to discrimination, both analytically and empirically (cf. Yates, 1982, 1994).

7. Concluding remarks

The present studies indicate that forecast consumers evaluate forecasting performance according to principles substantially different from the ones implicit in standard analytic methods. All of the distinctive features of those principles are interesting. However, we find three of them to be especially compelling and worthy of priority in future research. First, there is the emphasis on categorical 'correctness'. Although a good case can be made for the value of probabilistic rather than deterministic forecasts, it is not apparent that the typical consumer appreciates the advantages. Second, there is the significance consumers appear to attach to forecast extremeness. Among other things, some consumers interpret extremeness as an indicator of forecaster competence, others the opposite. Interestingly, the former interpretation suggests yet another possible contributor to the commonly observed phenomenon of overconfidence in probability judgments (cf. Yates, 1990, ch. 4); some people might exhibit extreme confidence because they (correctly?) believe that others will be impressed by such exhibitions. And then there is the seemingly very strong desire by consumers that forecasters be able to give good explanations for their predictions.

The practical implications of conclusions like those indicated here are considerable, assuming that they generalize. First of all, it is not at all clear that the approaches consumers take in evaluating forecasters are in the consumers' best interests. Indeed, as suggested at several points in this paper, there is reason to expect otherwise. An example is the insistence on plausible explanations for forecasts as opposed to statistically reliable demonstrations of accuracy. At the same time, consumers' multifaceted criteria for good forecasting suggest that current analytical methods for evaluating forecasts might be deficient in their narrowness. Moreover, it is clearly in the interests of practicing forecasters to be sensitive to the considerations consumers take into account. After all, a convincing forecaster, regardless of his or her 'objective' accuracy, is likely to be an employed forecaster.

Acknowlegements

This research was supported by U.S. National Science Foundation grant number SES92-10027 to the University of Michigan and by R.O.C. National Science Council grant number NSC94-2413-H033-002 to Chung Yuan University.

The authors thank the members of the University of Michigan's Judgment and Decision Laboratory for their suggestions about various aspects of this project. Special thanks to Jonathan Emmett, LeAnn Franke, and Winston Sieck who assisted with the collection and coding of data. We also appreciate the insightful com-

ments of Gideon Keren and two anonymous referees on an earlier version of this paper.

References

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

Buchanan, B.G. and E.H. Shortliffe, eds., 1984, *Rule-based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project* (Addison-Wesley, Reading MA).

Clemen, R.T., 1991, *Making Hard Decisions: An Introduction to Decision Analysis* (PWS-Kent, Boston, MA).

Cooper, W.H., 1981, Ubiquitous halo, *Psychological Bulletin*, 90, 218–244.

Jensen, F.A. and C.R. Peterson, 1973, Psychological effects of proper scoring rules, *Organizational Behavior and Human Performance*, 9, 307–317.

McGraw, B., 1990, Verdict boosts costs of Cobo by millions: City could have settled for $180,000, *Detroit Free Press*, 12 July, pp. 1A, 14A.

McNemar, Q., 1969, *Psychological Statistics* (Wiley, New York).

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

Murphy, A.H. and R.L. Winkler, 1992, Diagnostic verification of probability forecasts, *International Journal of Forecasting*, 7, 435–455.

Raiffa, H., 1968, *Decision Analysis* (Addison-Wesley, Reading, MA).

Shafer, G., 1976, *A Mathematical Theory of Evidence* (Princeton University Press, Princeton, NJ).

Siegel-Jacobs, K. and J.F. Yates, 1995, Effects of procedural and outcome accountability on judgment quality, *Organizational Behavior and Human Decision Processes*, in press.

Simonson, I., 1989, Choice based on reasons: The case of attraction and compromise effects, *Journal of Consumer Research*, 16, 158–174.

Slovic, P., D. Fleissner and W.S. Bauman, 1972, Analyzing the use of information in investment decision making, *Journal of Business*, 45, 283–301.

Teigen, K.H., 1990, To be convincing or to be right: A question of preciseness, in: K.J. Gilhooly, M.T.G. Keane, R.H. Logie and G. Erdos, eds., *Lines of Thinking: Reflections on the Psychology of Insight*, vol. 1 (Wiley, Chichester), pp. 299–313.

Tetlock, P.E., 1985, Accountability: A social check on the fundamental attribution error, *Social Psychology Quarterly*, 48, 227–236.

Weinstein, M.C. and H.V. Fineberg, 1980, *Clinical Decision Analysis* (Saunders, Philadelphia).

Winkler, R.L., 1969, Scoring rules and the evaluation of probability assessors, *Journal of the American Statistical Association*, 64, 1073–1078.

Winkler, R.L. and A.H. Murphy, 1968, 'Good' probability assessors, *Journal of Applied Meteorology*, 7, 751–758.

Yaniv, I., J.F. Yates and J.E.K. Smith, 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* (Prentice-Hall,Englewood Cliffs, NJ).

Yates, J.F., 1994, Subjective probability accuracy analysis, in: G. Wright and P. Ayton, eds., *Subjective Probability* (Wiley, Chichester), pp. 381–410.

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

Biographies: J. Frank YATES is Professor of Psychology at the University of Michigan. His research interests span various aspects of judgment and decision behavior, including basic processes, judgment accuracy analysis, cross-national variations, and decision aiding. He is the author of *Judgment and Decision Making*. He is also the associate editor of the *Journal of Behavioral Decision Making*, a consulting editor of *Psychological Review*, and a member of the editorial board of *Organizational Behavior and Human Decision Processes*.

Paul C. PRICE received his Ph.D. in psychology from the University of Michigan, where he is now a lecturer. His research interests include judgmental forecasting and confidence. Dr. Price's work has been published in the *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *Memory & Cognition*, and *Applied Cognitive Psychology*.

Ju-Whei LEE is Associate Professor in the Department of Psychology at Chung Yuan University. She received her Ph.D. from the University of Michigan. Her major research interests are in judgment and decision-making processes. Her articles appear in *Psychological Bulletin*, *Journal of Experimental Psychology: Human Performance and Perception*, *Philippine Journal of Internal Medicine*, and the *Asian Journal of Psychology*.

James RAMIREZ received his B.S. in psychology from the University of Michigan. He is currently a medical student at Wayne State University.