Test (1996)
Vol. **5**, *No.* 1, *pp.* 1–60

Scoring Rules and the Evaluation of Probabilities

R. L. WINKLER

Fuqua School of Business and Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708-0120, USA

[Read before the Spanish Statistical Society at a meeting organized by the Universitat de València on Tuesday, April 23, 1996]

SUMMARY

In Bayesian inference and decision analysis, inferences and predictions are inherently probabilistic in nature. Scoring rules, which involve the computation of a score based on probability forecasts and what actually occurs, can be used to evaluate probabilities and to provide appropriate incentives for "good" probabilities. This paper reviews scoring rules and some related measures for evaluating probabilities, including decompositions of scoring rules and attributes of "goodness" of probabilities, comparability of scores, and the design of scoring rules for specific inferential and decision-making problems.

Keywords: ATTRIBUTES OF "GOOD" PROBABILITIES; DECOMPOSITION OF EXPECTED SCORES; EVALUATION OF PROBABILITIES; PROBABILITY ASSESSMENT; PROBABILITY FORECASTS; SCORING RULES.

1. INTRODUCTION

In the complex world in which we live, we constantly have to make inferences and decisions in the face of uncertainty. Since probability is the mathematical language of uncertainty, it is natural in modeling inferential and decision-making problems to represent our uncertainty in terms of probabilities. The Bayesian approach to statistics formalizes

Received March 96; Revised April 96.

this modeling spirit, and the increasing application of Bayesian techniques means that more and more inferences and predictions are being expressed in probabilistic form. Furthermore, even when a full-blown Bayesian model is not constructed, probabilities are being assessed and used frequently. For example, precipitation probabilities are disseminated to the public regularly by the National Weather Service in the U.S. and several other countries (e.g., see Murphy and Winkler, 1984). Probabilities are assessed from experts or generated by models as a standard part of decision analysis (Raiffa, 1968; Howard and Matheson, 1983; Clemen, 1996) and risk analysis (Morgan and Henrion, 1990; Cooke, 1991). Probability forecasts are widely encountered in other areas as well, from economics ("there is a 40 percent chance that interest rates will rise this year") to medicine ("the probability that the patient will die during surgery is 0.1") to environmental issues ("the chance of this water supply being contaminated because of the waste disposal site nearby is one in 1,000") to international politics ("the chances are one in three that a treaty will be negotiated successfully by July").

Because they explicitly recognize and quantify uncertainty, probability forecasts are more informative and more valuable to decision makers than categorical forecasts ("interest rates will not rise this year") or forecasts with qualitative indications of uncertainty ("the water supply is unlikely to be contaminated"). To be able to measure how informative and valuable specific probability forecasts are, we need to evaluate them in some manner. The measures typically used for such evaluations include scoring rules, which involve the computation of a score based on the probability forecast and on the event (or value of the uncertain quantity) that actually occurs. In an ex ante sense, strictly proper scoring rules provide an incentive for careful and honest forecasting by the forecaster or forecast system. In an ex post sense, they reward accurate forecasts and penalize inferior forecasts. For example, if two forecasters provide probabilities of rain of 0.2 and 0.7, and it turns out that rain occurs, the higher probability receives the better score.

The purpose of this paper is to provide a review of scoring rules and some related measures for evaluating probabilities. The distinction between the roles of scoring rules in probability assessment (an ex ante incentive role) and probability evaluation (an ex post role) is discussed in Section 2. Some strictly proper scoring rules are reviewed in Section 3.

Scoring rules provide overall measures of "goodness" of probabilities, and Section 4 looks at measures of specific attributes of "goodness" (e.g., calibration and sharpness) that can be generated from decompositions of scoring rules. In Section 5, I discuss the comparability of scores. Section 6 addresses the connection of scoring rules to decision making and some broader design issues. Some brief concluding comments are presented in Section 7.

2. SCORING RULES IN PROBABILITY ASSESSMENT AND EVALUATION

The most natural role for scoring rules is simply to provide summary measures to evaluate probabilities in light of what actually happens. This is an ex post evaluation role in that it is conducted after we find out what actually happens. The development of scoring rules was also motivated by the desire to provide appropriate incentives for forecasters to honestly report their probabilities, as noted by de Finetti (1962, p. 359):

> The scoring rule is constructed according to the basic idea that the resulting device should oblige each participant to express his true feelings, because any departure from his own personal probability results in a dimunition of his own average score as he sees it.

This is an ex ante role that is part of probability assessment. Of course, the two roles are related, since the ex ante incentive is created by the anticipation of the ex post evaluation. Nonetheless, there are some important distinctions between the two roles.

As noted above, the ex ante role of scoring rules as incentives was motivated by subjective probability. The axioms underlying the theory of subjective probability (de Finetti, 1937; Savage, 1954) prescribe that probability assessors should be coherent in the sense that their subjective probabilities should obey the usual rules of probability theory. Any inconsistencies with these rules can and should be removed to yield coherent probabilities. But coherence is not sufficient to guarantee "good" probabilities. A forecaster assessing the probability of rain and the probability of no rain for tomorrow is coherent as long as the two probabilities add to one; coherence by itself provides no incentive for careful and honest probability assessment. This sort of concern has caused many to question the use of subjective probabilities and, by extension, the use of Bayesian methods, which involve the assessment of prior distributions.

A formal probability assessment process involves much more than incentives, of course. Important concerns include selecting appropriate experts to serve as probability assessors, gathering relevant past data and other information, training the experts in probability assessment, trying to counteract commonly found cognitive biases and simplifying heuristics, carefully defining the events and variables of interest, designing the actual probability assessment questions, and conducting the probability assessment itself. For discussions of general probability assessment issues, see Winkler (1967a), Spetzler and Staël von Holstein (1975), Wallsten and Budescu (1983), and Keeney and von Winterfeldt (1991).

A scoring rule can be any function of the probabilities and the observations. The ex ante incentive aspect of scoring rules, however, suggests that we should restrict our attention to strictly proper scoring rules, for which the assessor can maximize his or her expected score only by reporting probabilities honestly. The use of such rules in probability assessment weakens concerns that assessed subjective probabilities are in some sense arbitrary. Fortunately, the restriction to strictly proper rules still leaves us with a very rich set of available rules from which to choose. Moreover, the incentive aspect of strictly proper scoring rules extends beyond encouraging assessors to report their probabilities honestly; they also provide incentives for careful assessment and for gathering information in an attempt to "improve" probabilities. For example, the desire to maximize the expected score from a strictly proper scoring rule motivates the assessor to give probabilities that are well calibrated and sharp (Winkler, 1986); these attributes will be discussed in Section 4.

Much of the early work on scoring rules focused on the ex ante role of scoring rules and on the characterization of strictly proper scoring rules (e.g., Good, 1952; McCarthy, 1956; de Finetti, 1962, 1965; Shuford, Albert, and Massengill, 1966; Winkler, 1967b; Savage, 1971). However, Brier (1950) predates these references with his development of the Brier score, a quadratic scoring rule, from the viewpoint of ex post evaluation of weather forecasts expressed in terms of probabilities. Scoring rules have received considerable attention in the meteorological literature over the years (e.g., Winkler and Murphy, 1968; Murphy and Daan, 1985; Murphy, 1993; Wilks, 1995).

The Brier score is simply the mean squared error of the probability forecasts, where the observations are zeroes and ones. This is a natural

extension of the use of mean squared errors and similar measures to evaluate point forecasts. Informally, squared error seems to be an intuitively appealing measure to evaluate performance. Moreover, it turns out that this measure is strictly proper, and the appealing ex ante properties of strictly proper scoring rules translate into analogous appealing properties from an ex post viewpoint. For example, as noted above, with strictly proper scoring rules the assessor must report honestly to maximize his or her expected score. Ex post, this means that all other things being equal, the assessor will earn a higher average score if he or she is perfectly calibrated.

For reasons discussed above related to ex ante considerations, scoring rules have been rather closely associated with subjective probability. From an ex post evaluation viewpoint, it is clear that the use of scoring rules to evaluate probabilities has no connection with the source of the probabilities. Indeed, probability forecasts in meteorology have been generated by numerical-statistical models (models utilizing statistical techniques applied to the output from numerical models of the atmosphere) as well as subjectively by weather forecasters, and scoring rules have been used to evaluate the probabilities and to compare model-generated probabilities with subjectively-assessed probabilities (e.g., see Murphy and Winkler, 1984). Bernardo and Bermúdez (1985) use scoring rules to evaluate probabilistic classification models and to select the variables to be included in such models. Variable selection in this example, although based on an ex post evaluation, can also be viewed from an ex ante perspective in the sense of trying to build a classification model to maximize future expected scores.

Of course, from a subjective-probability viewpoint, all probabilities are subjective. When model-generated probability forecasts are used, they can be thought of as the subjective probabilities of the user or of the model builder. For consistency with this subjective-probability viewpoint and for convenience, I will write in terms of probability assessors generating subjective probabilities, but the underlying ideas and results also apply to model-generated probabilities.

Although some ex ante considerations with respect to scoring rules have their ex post counterparts, and vice versa, there are some distinctions. At the time the probabilities are assessed, all possible events or values of the variable of interest are relevant, and the assessor's expected

score depends on the entire probability distribution. After the relevant events or variables have been observed, it is natural to focus attention on the observed values and the probabilities assessed for those observed values. For instance, suppose that someone assesses probabilities r_1, r_2, and $1 - r_1 - r_2$ for the events "home team wins," "tie," and "home team loses" before a soccer match. Before the match, the assessor's expected score depends on both r_1 and r_2. The match winds up in a tie. After the match, should the evaluation of the probabilities depend only on r_2, or should it depend on both r_1 and r_2?

Most strictly proper scoring rules depend on both r_1 and r_2 because they give scores that depend not just on the probabilities assigned for observed values but on probabilities assigned for values that did not occur. However, such rules are inconsistent with Bayesian or other likelihood-based inferential procedures that can be used, for example, to evaluate alternative models generating probabilities in a given situation (e.g., see Roberts, 1965; Winkler, 1969; Bayarri and DeGroot, 1988). The likelihood principle indicates that only probabilities assigned to the observed values are relevant for inferential purposes. If we insist that our scoring rule be strictly proper and consistent with the likelihood principle, then for any situation with more than two possible outcomes, we must use a logarithmic scoring rule (Shuford, Albert, and Massengill, 1966). Considerations such as this are relevant in the choice of a scoring rule to use in a given situation, as are other design issues discussed in Section 6.

In the remainder of this paper, I will distinguish between probability assessment issues and probability evaluation issues only when the distinction is important. To avoid confusion, ex ante notation with expected scores will be used when characteristics of scoring rules are discussed.

3. STRICTLY PROPER SCORING RULES

In this section, some strictly proper scoring rules for various situations are reviewed. First, consider a simple dichotomous situation involving a single event A and its complement. Since coherence is assumed, we can restrict our attention to the probability of A. Let p denote the assessor's subjective probability that A will occur, and let r denote the probability of A actually reported by the assessor. A scoring rule S gives the assessor a score $S_1(r)$ if A occurs and $S_2(r)$ if A does not occur. The assessor's

expected score is then

$$E_p[S(r)] = pS_1(r) + (1-p)S_2(r), \tag{1}$$

and S is said to be strictly proper if

$$E_p[S(p)] > E_p[S(r)] \qquad \text{for} \qquad r \neq p. \tag{2}$$

Thus, to maximize expected score, the assessor should set $r = p$ (i.e., the assessor should be honest).

Three frequently-used scoring rules for the single-event situation are the quadratic, logarithmic, and spherical rules:

$$\text{Quadratic:} \quad S_1(r) = -(1-r)^2, \qquad S_2(r) = -r^2, \tag{3}$$

$$\text{Logarithmic:} \quad S_1(r) = \log r, \qquad S_2(r) = \log(1-r), \tag{4}$$

$$\text{Spherical:} \quad S_1(r) = r/[r^2 + (1-r)^2]^{1/2},$$

$$S_2(r) = (1-r)/[r^2 + (1-r)^2]^{1/2}. \tag{5}$$

Sometimes different variants of these rules are encountered, since any positive linear transformation of a strictly proper scoring rule is itself strictly proper. Also, some scores have a negative orientation, which means that a lower score is better and the goal is to minimize expected score; the Brier score (Brier, 1950) is simply twice the negative of the quadratic score given in (3).

Of course, the rules given in (3)-(5) are but three examples of an infinity of possible strictly proper rules. Savage (1971) and Schervish (1989) give two different representations of the class of all strictly proper scoring rules. I find Savage's representation easier to relate to some attributes of scoring rules that will be discussed in Sections 4 and 5; this is because it relates to the expected-score function $E_p[S(p)] = E_r[S(r)]$ for an assessor who sets $r = p$ and thereby maximizes his or her expected score. Savage's representation can be stated quite simply: A strictly proper scoring rule can be generated from any strictly convex expected-score function. For the single-event situation, the expected-score function is defined on the unit interval, and we find that

$$S_1(r) = E_r[S(r)] + (1-r)d\{E_r[S(r)]\}/dr \tag{6}$$

and

$$S_2(r) = E_r[S(r)] - rd\{E_r[S(r)]\}/dr. \tag{7}$$

The concept of a scoring rule can be generalized to a situation involving more than a single event and its complement. Consider a set of mutually exclusive and exhaustive events $\{A_i | i \in I\}$, where I is a finite or countably infinite set. Let p_i denote the assessor's probability that A_i will occur, let r_i denote the probability of A_i actually reported by the assessor, and define $p = (p_1, p_2, ...)$ and $r = (r_1, r_2, ...)$. In this paper, p and r represent single probabilities when a single event is of interest, vectors of probabilities when several events are of interest, and mass or density functions when a random variable is of interest. This abuse of notation is used to emphasize the distinction between an assessor's probability judgments (p) and reported probability judgments (r).

If a scoring rule gives a score $S(r) = S_j(r)$ if A_j occurs, then the assessor's expected score is

$$E_p[S(r)] = \sum_{j \in I} p_j S_j(r). \tag{8}$$

The scoring rule is strictly proper if (2) is satisfied. Quadratic, logarithmic, and spherical rules in this situation are, respectively,

$$S_j(r) = 2r_j - \sum_{i \in I} r_i^2, \tag{9}$$

$$S_j(r) = \log r_j, \tag{10}$$

and

$$S_j(r) = \frac{r_j}{\left(\sum_{i \in I} r_i^2\right)^{1/2}}. \tag{11}$$

Scoring rules for continuous random variables can also be developed. Consider a continuous random variable y. Let p denote the assessor's probability density function for y, and let r denote the density function reported by the assessor. If a scoring rule gives a score $S_x(r)$ when $y = x$, then the assessor's expected score is

$$E_p[S(r)] = \int_{-\infty}^{\infty} S_x(r) p(x) dx. \tag{12}$$

The scoring rule is strictly proper if (2) is satisfied. Quadratic, logarithmic, and spherical rules in this situation are, respectively,

$$S_x(r) = 2r(x) - \int_{-\infty}^{\infty} r^2(x) dx, \tag{13}$$

$$S_x(r) = \log r(x), \tag{14}$$

and

$$S_x(r) = \frac{r(x)}{\left(\int_{-\infty}^{\infty} r^2(x)dx\right)^{1/2}}. \tag{15}$$

The expected score $E_p[S(r)]$ refers to a single occasion for which the assessor has probability (or probability distribution) p and reports probability (or probability distribution) r. It is an expected score given p. We can look at the expected-score function to see how the assessor's expected score varies as p varies or how the expected scores of two or more assessors differ when the assessors do not have the same p.

Next, let us step back and consider the expected score before p is known. If p is not known but has probability density f, the expected score is

$$E_f\{E_p[S(r)]\} = \int E_p[S(r)]f(p)dp. \tag{16}$$

Here f could be interpreted from an analyst's point of view as representing the analyst's distribution of the assessor's p, or it could be interpreted from the assessor's point of view as representing the assessor's distribution of p before the assessor has taken the time to look at the relevant information and think carefully about the situation. For example, imagine a weather forecaster contemplating potential probabilities of precipitation (to be assessed later, only one day in advance) for a day that is a month in the future or a physician contemplating potential probabilities of survival for the next patient before knowing anything about that patient. For those who prefer to think in terms of a series of trials, f could represent a distribution of probabilities over a long series of exchangeable trials.

The above notation and discussion refer to an ex ante perspective, which will be maintained throughout this paper. However, it might be helpful to clarify the connection between ex ante and ex post perspectives. Over a series of occasions with the same value r given as the probability forecast for a single-event situation by an assessor, the average score ex post is

$$\overline{S}_q(r) = qS_1(r) + (1-q)S_2(r), \tag{17}$$

where q is the relative frequency of the event on those occasions. For example, we could consider all of the days on which a weather forecaster

reports a probability of precipitation of $r = 0.3$ and find that precipitation actually occurs on 18% of those days (i.e., $q = 0.18$). Comparing (1) and (17), we see that

$$\overline{S}_q(r) = E_q[S(r)]. \tag{18}$$

Similarly, if we look at a series of occasions on which different values of r may be given, then the overall average score ex post is

$$\overline{S} = E_g[\overline{S}_q(r)] = E_g\{E_q[S(r)]\}, \tag{19}$$

where g represents the observed relative frequency distribution of r. Thus, the ex post average scores are of the same form as the ex ante expected scores, with relative frequencies q and g replacing probabilities p and f.

4. DECOMPOSITIONS OF SCORING RULES: ATTRIBUTES OF "GOOD" PROBABILITIES

I like to think of scoring rules as encouraging honesty and careful assessment and as providing overall measures of "goodness" of probability forecasts. But in addition to an overall evaluation, we can look at specific attributes of probabilities, and scoring rules can be decomposed to provide measures relating to these attributes. The study of decompositions dates back at least to the decompositions of the Brier score given by Sanders (1963). Other useful references in this arena include Murphy (1972a, 1972b, 1973b), DeGroot and Fienberg (1982, 1983), Yates (1982, 1988), DeGroot and Eriksson (1985), Yates and Curley (1985), Blattenberger and Lad (1985), and Murphy and Winkler (1987, 1992).

For a strictly proper scoring rule, the expected score $E_p[S(r)]$ can be expressed as

$$E_p[S(r)] = E_p[S(p)] + C(S, r, p), \tag{20}$$

where $C(S, r, p)$ is maximized (at zero) when $r = p$. This follows directly from (2). In an ex ante sense, $C(S, r, p)$ is a penalty for any deviation of r from p. The condition that $r = p$ can be thought of as honest reporting, but it can also be thought of as ex ante calibration as viewed by the assessor. As noted by Dawid (1982), a coherent assessor expects to be calibrated. In an ex post sense, the relative frequency q

replaces p, and C can then be viewed as a penalty for being miscalibrated in a frequency sense.

Calibration is one attribute of "goodness" of probability forecasts, and (20) provides a decomposition which expresses the expected score as a sum of the expected score for a perfectly-calibrated assessor and the penalty for miscalibration. From Savage (1971), $E_p[S(p)]$ is a strictly convex function, and for most scoring rules used in practice (including the quadratic, logarithmic, and spherical rules given in Section 3), it is a symmetric function. For the single- event situation, this symmetry holds when $S_1(u) = S_2(1-u)$ for all $u \in [0,1]$ and means that $E_p[S(p)]$ is minimized at $p = 0.5$ and maximized at zero or one.

Thus, holding calibration constant, assessors who are able to give more extreme forecasts have higher expected scores. The "holding calibration constant" prevents assessors from being able to benefit by arbitrarily giving extreme probabilities even when their values of p are not extreme; note that $E_p[S(p)]$ is a function of p, not a function of r. As a result, although the expected score is expressed in (20) as a sum of terms rewarding extremeness and calibration, it is not the case that the assessor can trade these two conditions off against each other by sacrificing calibration to get a better score on extremeness.

As in Section 3, we can consider the expected score before p is known. From (16) and (20),

$$\begin{aligned} E_f\{E_p[S(r)]\} &= \int E_p[S(p)]f(p)dp + \int C(S,r,p)f(p)dp \\ &= E_f\{E_p[S(p)]\} + E_f[C(S,r,p)]. \end{aligned} \tag{21}$$

The first term on the R.H.S. of (21) is the overall expected score for a perfectly-calibrated assessor, and it rewards assessors who are more likely to have extreme probabilities and less likely to have probabilities near one-half. This characteristic is often referred to as sharpness, or refinement. Of course, the specific way in which refinement is rewarded will vary for different strictly proper scoring rules. DeGroot and Fienberg (1982, 1983) give a formal definition of refinement that provides a partial order of perfectly-calibrated assessors independent of the specific scoring rule being used.

For the quadratic scoring rule given in (3),

$$E_f\{E_p[S(r)]\} = -\int p(1-p)f(p)dp - \int (r_p - p)^2 f(p)dp, \tag{22}$$

where r_p represents the reported probability r as a function of p. The first term indicates that sharpness is measured by $-E_f[V(x|p)]$, where $x = 1$ if A occurs and $x = 0$ if A does not occur. The second term shows that deviations from perfect calibration are penalized via the expected squared difference between r and p.

Strictly proper scoring rules such as the quadratic rule can be decomposed in different ways, yielding measures of other attributes of probability forecasts. These attributes include, for example, discrimination (the degree to which the probabilities discriminate between occasions on which A occurs and occasions on which A does not occur), resolution [the degree to which conditional means $E(x|p)$ differ from the base rate $E(x)$], and unconditional bias [the degree to which the overall mean probability $E(p)$ differs from the base rate]. For more details on attributes of probability forecasts and decompositions of scoring rules, see Murphy and Winkler (1987, 1992) and Murphy (1993). These papers also present an approach for evaluating probability forecasts that is based on the joint distribution of forecasts and observations and the marginal and conditional distributions that can be obtained from this joint distribution. The joint-distribution approach and scoring rules (with their decompositions) provide a rich framework for evaluating probability forecasts in a diagnostic sense.

5. THE COMPARABILITY OF SCORES

As discussed in Section 4, strictly proper scoring rules encourage calibration and reward sharpness. This provides appropriate incentives from an ex ante viewpoint, but how can we interpret the numerical average scores that are computed ex post? Consider, for example, the quadratic scoring rule given by (3) for the single-event situation. The best possible score is zero for forecasts of zero and one which are never wrong (i.e., perfect forecasts), and the worst possible score is -1 for perfect forecasts of zero and one which are always wrong. The expected score for a well-calibrated forecaster ranges from -0.25 (when $p = 0.5$) to zero (when $p = 0$ or 1). Suppose we use this score to evaluate a weather forecaster who has made a long series of precipitation probability forecasts, and the average score is -0.08. What can we say about the performance of this forecaster?

Unfortunately, the scores obtained by a forecaster will depend not just on his or her forecasting ability, but on the nature of the forecasting situations he or she faces. Based on samples of roughly 10,000 precipitation probability forecasts and observations at different locations in the U.S. (Winkler, 1994), the forecaster with the average score of -0.08 would be doing quite well in Portland, Oregon (where the average quadratic score for the sample period was -0.11) but not so well in Phoenix, Arizona (average score of -0.04). In a location where climatology (the historical relative frequency of precipitation) is 0.09 (as is the case in Tucson, Arizona, for instance), a forecaster could expect to get an average score of -0.08 by simply reporting an r equal to climatology each day.

It seems, then, that attempting to evaluate probabilities in an absolute sense is a tricky business. Might we be better off to think of evaluation in a comparative sense? Perhaps, but comparative in what sense? We might compare one weather forecaster to another. If the two forecasters assess probabilities for exactly the same series of situations, then their scores would seem to be comparable. However, different forecasters generally do not deal with the same situations; they may provide forecasts on different days or for different locations, for example.

The development of so-called "skill scores" (see Murphy, 1973a, 1974, 1996) has been motivated by the desire to provide average scores that reflect the relative ability of a forecaster rather than some combination of the forecaster's ability and the situation's difficulty. Skill scores use a strictly proper scoring rule and compare a forecaster's average score to the average score that some sort of benchmark forecasting rule would have obtained for the same set of situations. Usually the benchmark is an unsophisticated rule such as climatology, which is just a base rate forecast. The most commonly used skill score is simply the percentage improvement over climatology in average score. If the strictly proper rule is S, then we have

$$\text{Skill} = \frac{S - S_{cl}}{S_{cl}}, \tag{23}$$

where S_{cl} is the average score for climatology. Although (23) may seem intuitively appealing, it is not strictly proper and its values will change with a linear transformation of S (which does not change the strictly proper nature of S).

An alternative approach that yields strictly proper scoring rules that are standardized in some sense and are capable of representing an evaluator's judgments about the difficulty of different forecasting situations is developed in Winkler (1994). The idea is to generate a strictly convex expected-score function to represent a forecasting situation and then to use (6) and (7) (or appropriate extensions for other than the single-event situation) to find a strictly proper scoring rule that yields the desired expected-score function. These rules are called asymmetric scoring rules because, in contrast to the usual rules such as the standard quadratic, logarithmic, and spherical rules given in Section 3, their expected-score functions are not generally symmetric.

For example, it seems reasonable to want $E_p[S(p)]$ to be high for $p = 0$ and $p = 1$ and to be lowest at a forecast judged to be "least skillful." Climatology is arguably least skillful among well-calibrated precipitation probability forecasts. Alternatively, the benchmark forecast receiving the lowest average score could be based on the utility of forecasts in a decision-making problem or on some other consideration; design issues such as this will be discussed in Section 6.

Denoting the benchmark least skillful forecast in the single-event situation by $c[c \in (0,1)]$, we can generate a family of strictly proper asymmetric scoring rules S^* corresponding to any strictly proper symmetric rule S:

$$S^*(r) = \frac{S(r) - S(c)}{T(c)}, \tag{24}$$

where $T(c) = S_1(1) - S_1(c)$ if $r \geq c$ and $T(c) = S_2(0) - S_2(c)$ if $r \leq c$ (Winkler, 1994). S^* is standardized in the sense that the average score $E_p[S(p)]$ has a minimum of zero at $p = c$ and a maximum value of one at $p = 1$ and $p = 0$. For example, if S is the quadratic rule defined by (3),

$$S_1^*(r) = \frac{(1-c)^2 - (1-r)^2}{T(c)} \quad \text{and} \quad S_2^*(r) = \frac{c^2 - r^2}{T(c)}, \tag{25}$$

where $T(c) = c^2$ if $r \leq c$ and $(1-c)^2$ if $r \geq c$. The expected-score function for an assessor who sets $r = p$ in accordance with the strictly proper nature of the rule is $E_p[S(p)] = (p-c)^2/T(c)$.

If directly comparable scores (scores from different forecasters for the same set of situations) are not available, rules such as the asymmetric

rules provide reasonable alternatives for relative evaluation. Moreover, the standardized nature of the rules suggests that in a rough sense, they might help a bit in the more difficult absolute evaluation. The discussion of design issues in Section 6 will touch on some issues relevant to the use of asymmetric rules.

6. WHICH SCORING RULE?

All strictly proper scoring rules encourage honesty and careful assessment ex ante and reward calibration and sharpness ex post. How, then, might we choose a specific rule to use in practice? Different scoring rules have different characteristics, as exemplified by the distinctions between symmetric scoring rules and the asymmetric rules discussed in Section 5. Also, different situations and goals may call for different types of rules. Sometimes we are interested simply in an overall evaluation of the assessor as an expert (e.g., as a weather forecaster), other times the assessor is secondary to our concern with a particular set of probabilities, and yet other times we are focusing primarily on a decision-making problem for which the assessor can provide useful information. There is no universal answer to the question posed in the title of this section, but an understanding of some characteristics of certain types of scoring rules can help in the choice of a rule, and rules can be tailored somewhat to fit particular situations. In this section, I briefly address some issues that seem relevant to the choice of a scoring rule.

Local Scoring Rules. When we go beyond a simple dichotomy, the assessor is asked to assess not just a single probability, but an entire probability distribution. Then we observe a single event or value of the variable of interest. A scoring rule is called local if the score depends only on the probability or density assigned to what is observed. For example, consider the scoring rules given in (9) - (11) and (13) - (15). The logarithmic rules in (10) and (14) are local, whereas the quadratic and spherical rules are not local because they depend on the entire probability distribution. In fact, the logarithmic scoring rule is the only strictly proper scoring rule that is local (Shuford, Albert, and Massengill, 1966) for other than the case of a simple dichotomy.

When might a local scoring rule be of interest? According to the likelihood principle, in an inferential problem the full import of an observation is contained in the likelihoods for that observation, and likelihoods

for other events or values that might have been observed but were not are irrelevant. In a Bernoulli experiment with parameter π, if we observe five successes in twenty trials, our inferences are based on $\pi^5(1-\pi)^{15}$, not on $\pi^x(1-\pi)^{20-x}$ for any value of x other than five. Thus, to be consistent with the likelihood principle in formal inferential evaluations of assessors (or of models generating probabilities), we need to use a logarithmic rule (Winkler, 1969), as noted in Section 2. This is discussed by Bernardo and Smith (1994, p. 72):

> It is intuitively clear that the preferencess of an individual scientist faced with a pure inferential problem should correspond to the ordering induced by a local score function. The reason for this is that, by definition, in a 'pure' inference problem we are solely concerned with 'the truth.' It is therefore natural that if E_j, say, turns out to be true, the individual scientist should be assessed (i.e., scored) only on the basis of his or her reported judgement about the plausibility of E_j.

Sensitivity to Distance. Suppose that three different assessors give probabilities for x, the number of games a basketball team wins in a four-game tournament. The possible values for x are 0, 1, 2, 3, and 4. The vectors of assessed probabilities for these five values are (0.1, 0.1, 0.6, 0.1, 0.1) for assessor A, (0, 0.2, 0.6, 0.2, 0) for assessor B, and (0.2, 0, 0.6, 0, 0.2) for assessor C. The tournament is conducted, and the team wins two games. All three assessed a probability of 0.6 for this observed value of x. With the quadratic scoring rule given in (9), A receives a score of 0.80, and B and C both receive scores of 0.76. This illustrates the fact that if a scoring rule is not local (i.e., it is not logarithmic), the scores it yields will depend on the probabilities for values other than the observed value. Moreover, note that a fourth assessor D who gives probabilities (0, 0, 0.65, 0.35, 0) receives a quadratic score of 0.755. This score is less than the scores received by the other three assessors, all of whom assigned a lower probability to what was actually observed (the team winning two games).

The example also illustrates the notion of sensitivity to distance, which involves the probability assigned to values close to the observed value as compared with probabilities assigned to values farther from the true value. Looking at the total probability of 0.4 assigned to values other than the observed value ($x = 2$) by A, B, and C, we see that B assigned all of this 0.4 to the values ($x = 1$ and $x = 3$) adjacent to $x = 2$. In contrast, C assigned all of the 0.4 to the values ($x = 0$ and

$x = 4$) farthest from $x = 2$. A is in between, with 0.2 assigned to the adjacent values and 0.2 assigned to the extreme values. In this sense C's probabilities for $x \neq 2$ are "more distant" from $x = 2$ than A's probabilities, which in turn are "more distant" from $x = 2$ than B's probabilities. A strictly proper scoring rule that is sensitive to distance, such as the ranked probability score (Epstein, 1969), would give B the highest score, followed by A and then C. The quadratic score, on the other hand, is not sensitive to distance and gives A the highest score. Matheson and Winkler (1976) develop scoring rules that are sensitive to distance in the continuous case.

If we believe that a set of probabilities less distant from the true value reflects greater skill on the part of the assessor, then a scoring rule that is sensitive to distance might be reasonable. Such might also be the case when the probabilities are being used as information in a decision-making problem. The initial development of scoring rules that are sensitive to distance (Epstein, 1969; Staël von Holstein, 1970) was based on a particular class of decision-making problems.

Adjusting Scoring Rules for Difficulty. Since a strictly proper scoring rule can be generated from any strictly convex expected-score function $E_p[S(p)]$, as noted in Section 3, the choice of a rule to use in a specific situation can be viewed as a choice of an expected-score function. This is the spirit underlying the development of the asymmetric scoring rules discussed in Section 5. In an attempt to adjust for difficulty and tailor the scoring rule to a particular situation, we can choose an expected-score function that minimizes the expected score at the probability judged to be the "least skillful" forecast. This least skillful forecast could be a base rate such as climatology, or it could be some other value. The expected score should be maximized at the "most skillful" forecast, which is a perfect forecast (e.g., a probability of zero or one in the assessment of a probability for a single event). For the quadratic asymmetric rule in (25), the expected score consists of two quadratic functions, one defined on $[0, c]$ and the other on $[c, 1]$, with a minimum value of zero at c and a maximum value of one at 0 and 1.

If one extreme forecast is viewed as more skillful (perhaps because it is more difficult to forecast precipitation perfectly than to forecast no precipitation perfectly, for example), then the expected score should be maximized at that extreme point. The value of the expected score at the

other extreme point can be assessed to correspond to the perceived degree of skill associated with that forecast. For example, with the quadratic rule given in (25), rescaling one of the two quadratic functions (either above or below c) so that it reaches a value of less than one at the extreme point is a simple generalization that enriches the family of asymmetric rules.

The spirit of the general development of measures to represent the skill of probability forecasts is very much like that of preference assessment in decision analysis (e.g., see Keeney and Raiffa, 1976). Minimum and maximum values of an expected-score function are chosen to correspond to the desired scaling (e.g., zero to one) and to occur at the probability values judged to represent the least and most skillful forecasts.

Scoring Rules and Decision Making. Adjusting scoring rules for difficulty amounts to tailoring scoring rules for difficulty in a forecasting sense. Often the raison d'être for the assessment of probabilities is to help a decision maker facing a decision-making problem under uncertainty. Scoring rules can be designed to be consistent with the utility of forecasts in decision-making problems. For a discussion of the notion of a scoring rule as a "share of a business" to provide appropriate motivation for careful and accurate forecasting, see Savage (1971). Pearl (1978, p. 176) notes that

> a scoring rule can be chosen which passes on to the forecaster some of the economical consequences of his report, as viewed by the client.

The value of various types of forecasts, including probability forecasts, in weather-related decision-making problems such as the cost-loss ratio situation has been investigated in some detail and related to scoring rules (e.g., see Murphy, 1977). A general method for comparing probability forecasts based on a loss function in a two-decision problem is developed and related to proper scoring rules in Schervish (1989).

When the design of a scoring rule is related to a decision-making problem, the emphasis shifts from the skill of the forecasts to the value of the forecasts in the context of the decision-making problem. Information has value in a decision-making problem to the extent to which it is likely to cause a change in the decision. Thus, for example, a forecast providing a well-calibrated probability that is the same as the decision maker's own prior probability is not valuable because it will not change

the preferred course of action. This suggests that it might be desirable to minimize the expected-score function at the decision maker's prior probability. Alternatively, in a simple two-action problem, a probability forecast equal to the breakeven probability (the probability for which the two actions yield identical expected utilities) is not valuable because it does not suggest preference for one decision over the other. Schervish's method for comparing probability forecasts relates to a breakeven probability and can be used to generate asymmetric rules (Schervish, 1989). Matheson and Winkler (1976) introduce a weighting function that can give greater weight to probabilities for certain regions of values of the variable of interest, thereby encouraging the assessor to pay more attention to probabilities in such regions.

The process of tailoring scoring rules to particular decision-making situations might be formalized further by attempting to relate the scoring rule to the expected value of information for the probability forecast. Also, there may be a feedback loop that ties the decision back to the probability forecast. For example, consider the elicitation of a sales manager's probabilities for next year's sales. We want to get an accurate picture of the prospects for next year, but we also want the incentive plan to inspire the manager to work hard to maximize sales. Sarin and Winkler (1980) develop a reward function that is a weighted average of a strictly proper scoring rule and a function that is increasing in sales and does not depend on the assessed probabilities. This reward function simultaneously encourages honest assessment ex ante and effort to maximize sales ex post.

It is possible, then, to relate scoring rules to decision making. The basic idea is to attempt to align the incentives for the assessor (and therefore the resulting evaluations as well) with the needs of the decision-making problem. Once we start tailoring scoring rules to individual decision-making problems and the value of the probability forecasts, of course, the resulting scores will be highly dependent on the specific problem and on the specific decision maker and may not be well-suited for more general-purpose evaluation.

Risk-Averse or Risk-Taking Probability Assessors. For the purpose of probability assessment, the notion of strictly proper scoring rules is based on the premise that the assessor wants to maximize his or her expected score. This implicitly assumes that the assessor's utility function

is linear in the score. (If the score is translated into a monetary payoff that is a linear function of the score, the assessor's utility function for money is relevant, but we can consider the utility of the score directly regardless of whether monetary payoffs are involved.) If the assessor is not risk neutral with respect to the score, then the incentives are skewed and the score will no longer promote honest reporting of probabilities.

In principle, adjustments for nonlinear utility are straightforward (Winkler, 1969). Suppose that the assessor's utility function for the score is known, and denote it by U. Assuming that U is strictly monotone, as would be anticipated, then it has an inverse, U^{-1}. If S is a strictly proper scoring rule under a linear utility function (i.e., it yields a maximum expected score if and only if the assessor sets $r = p$), then $U^{-1}(S)$ is a strictly proper scoring rule in a utility sense (i.e., it yields a maximum expected utility if and only if the assessor sets $r = p$). For example, suppose that S is the logarithmic scoring rule given in (4) and that the assessor's utility function for the score is itself logarithmic, $U(S) = \log S$. The composite function $U^{-1}(S)$ is simply a linear function with $S_1(r) = r$ and $S_2(r) = 1 - r$. Therefore, a linear scoring rule, which is not strictly proper in the sense of maximizing the expected score (it encourages a risk-neutral assessor to set $r = 1$ when $p > 0.5$ and $r = 0$ when $p < 0.5$), does encourage honest assessment for an assessor with a logarithmic utility function.

Of course, this adjustment for nonlinear utility assumes knowledge of the assessor's utility function. Since we generally will not know U or want to expend the time and effort required to elicit it from the assessor, any adjustments will have to be based on our judgments about U. Also, even if we do know U, another fly in the ointment is the possible presence of other stakes that the assessor may have in the events or variables of interest. Kadane and Winkler (1988) show how such other stakes can cause systematic shifts in assessed probabilities if the assessor's utility function is nonlinear.

Feedback and Learning. The discussion in this paper has focused on the ex ante incentive concerns and ex post evaluation aspects of scoring rules with respect to a particular probability forecast or set of probability forecasts. Also of interest are the implications for future probability forecasts beyond those being made and evaluated. In this sense, scores can provide useful feedback to give an assessor information

about his or her performance. This information, of course, can also be helpful to an analyst or decision maker. Since different scoring rules have different characteristics, multiple scoring rules might be used for feedback purposes.

Scores from one or more strictly proper scoring rules are summary measures of performance. Looking beyond these summary measures can facilitate effective learning and improvement on the part of the assessor. As noted in Section 4, decompositions of scoring rules enable us to look at specific attributes of probabilities such as calibration and sharpness. For instance, two assessors providing probability forecasts for the same set of situations could receive identical average scores even though the characteristics of their probabilities are very different. The probabilities from the first assessor might be very well-calibrated but not very sharp, while those from the second assessor might be quite sharp but poorly calibrated. This suggests that the first assessor should work on trying to be able to provide more extreme probabilities without sacrificing calibration and the second assessor should strive to become better calibrated. Of course, this feedback and learning process can be just as helpful in the development and refinement of models that generate probability forecasts as it is for assessors who provide subjective probabilities.

The extra information provided by decompositions above and beyond overall expected scores suggests that for feedback and learning purposes, we might want to select scoring rules for which convenient and well-understood decompositions are available. Investigations of decompositions have concentrated on quadratic rules, perhaps because they are the most commonly-used scoring rules in practice and because they are easy to decompose. An exploration of decompositions of other rules would increase our storehouse of measures of various attributes of "goodness" of probabilities. Moreover, feedback other than scoring-rule feedback can be of particular value in helping assessors learn and improve as well as in helping analysts or decision makers evaluate various characteristics of different assessors. The diagnostic approach based on the joint distribution of probabilities and observations (Murphy and Winkler, 1987, 1992) is very promising in this regard. For an application of this diagnostic approach in the context of physicians assessing proba-

bilities of survival for patients in an intensive care unit, see Winkler and Poses (1993).

7. CONCLUDING COMMENTS

In this paper, I have attempted to review the development and use of scoring rules and to highlight some important issues related to such rules, skipping over technical details to emphasize the key concepts and ideas. It is only natural that such a review is personal, slanted toward some of my own work and my own particular interests and biases. I welcome the thoughts of others (agreeing or disagreeing) on these issues. No attempt has been made to be exhaustive in the listing of references, but they should be sufficient to provide useful signposts for those wanting to familiarize themselves with the literature. It is a diverse literature; although the development and study of scoring rules has strong roots in statistics (building on the ideas of de Finetti, Savage, and others), important contributions have come from meteorology (ranging from the early work of Brier to the extensive contributions by Murphy), decision analysis, and psychology. Different fields bring different perspectives to the table: statisticians may tend to focus more on inferential problems and general-purpose incentives and evaluations, meteorologists on evaluation and measures of skill, decision analysts on the relation of scoring rules to decision-making problems (no surprise there), and psychologists on performance on specific attributes (especially calibration). But this is a gross oversimplification, and there has been a great deal of cross-fertilization.

At the beginning of this paper, I noted that we constantly have to make inferences and decisions in the face of uncertainty in the complex world in which we live. The advancement of the Bayesian approach has increased the degree to which inferences and predictions are expressed in probabilistic form in statistical applications, decision analysis has furthered the use of probabilities in the modeling of decision-making problems, and probabilistic risk analysis has made inroads lately. Of course, many of these probabilities are not evaluated; in some cases, ex post evaluation is impossible within any reasonable time frame (e.g., consider the probability that the mean temperature at a prospective nuclear waste disposal site will increase by at least three degrees Celsius in the next 1,000 years) or because of the nature of the probabilities (e.g.,

probabilities for unobservable parameters of models as opposed to probabilities for observable events or variables). Furthermore, probability receives short shrift in our educational system, and exposure to probability is unduly limited in everyday life. Thus, even though probabilities are being assessed and used frequently, the potential for greater use of probabilities and scoring rules is almost limitless. Further work on the concepts and ideas discussed in this paper and applications of scoring rules and related probability-evaluation measures would be most helpful in this regard.

REFERENCES

Bayarri, M. J. and DeGroot, M. H. (1988). Gaining weight. A Bayesian approach. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 25–44 (with discussion).

Bernardo, J. M. and Bermúdez, J. D. (1985). The choice of variables in probabilistic classification. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 67–81 (with discussion).

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley

Blattenberger, G. and Lad, F. (1985). Separating the Brier score into calibration and refinement components: A graphical exposition. *Amer. Statist.* **39**, 26–32.

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

Clemen, R. T. (1996). *Making Hard Decisions*, 2nd Edition. Belmont, CA: Duxbury Press.

Cooke, R. M. (1991). *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford: University Press.

Dawid, A. P. (1982). The well-calibrated Bayesian. *J. Amer. Statist. Assoc.* **77**, 605–613.

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré* **7**, 1–68. Translated as "Foresight: Its logical laws, its subjective sources" in *Studies in Subjective Probability* (H. E. Kyburg and H. E. Smokler, eds.). New York: Wiley, 1964, 93–158.

de Finetti, B. (1962). Does it make sense to speak of "good probability appraisers"? *The Scientist Speculates: An Anthology of Partly-Baked Ideas* (I. J. Good, ed.). New York: Wiley, 357–363.

de Finetti, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item. *British J. of Math. and Stat. Psych.* **18**, 87–123.

DeGroot, M. H. and Eriksson, E. A. (1985). Probability forecasting, stochastic dominance, and the Lorenz curve. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 99–118 (with discussion).

DeGroot, M. H. and Fienberg, S. E. (1982). Assessing probability assessors: Calibration and refinement. *Statistical Decision Theory and Related Topics III* **1** (S. S. Gupta and J. O. Berger, eds.). New York: Academic Press, 291–314.

DeGroot, M. H. and Fienberg, S. E. (1983). The comparison and evaluation of forecasters. *The Statistician* **32**, 14–22.

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *J. Appl. Meteorology* **8**, 985–987.

Good, I. J. (1952). Rational decisions. *J. Roy. Statist. Soc. B* **11**, 107–114.

Howard, R. A. and Matheson, J. E. (eds.) (1983). *The Principles and Applications of Decision Analysis* (2 volumes). Palo Alto, CA: Strategic Decisions Group.

Kadane, J. B. and Winkler, R. L. (1988). Separating probability elicitation from utilities. *J. Amer. Statist. Assoc.* **83**, 357–363.

Keeney, R. L. and Raiffa, H. (1976). *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. New York: Wiley.

Keeney, R. L. and von Winterfeldt, D. (1991). Eliciting probabilities from experts in complex technical problems. *IEEE Trans. Eng. Management* **38**, 191–201.

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Manag. Sci.* **22**, 1087–1096.

McCarthy, J. (1956). Measures of the value of information. *Proc. Nat. Acad. Sciences* **42**, 654–655.

Morgan, M. G. and Henrion, M. (1990). *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge: University Press.

Murphy, A. H. (1972a). Scalar and vector partitions of the probability score. Part I. Two-state situation. *J. Appl. Meteorology* **11**, 273–282.

Murphy, A. H. (1972b). Scalar and vector partitions of the probability score. Part II. N-state situation. *J. Appl. Meteorology* **11**, 1183–1192.

Murphy, A. H. (1973a). Hedging and skill scores for probability forecasts. *J. Appl. Meteorology* **12**, 215–223.

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

Murphy, A. H. (1974). A sample skill score for probability forecasts. *Monthly Weather Review* **102**, 48–55.

Murphy, A. H. (1977). The value of climatological, categorical, and probabilistic forecasts in the cost-loss ratio situation. *Monthly Weather Review* **105**, 803–816.

Murphy, A. H. (1993). What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting* **8**, 281–293.

Murphy, A. H. (1996). General decompositions of MSE-based skill scores: Measures of some basic aspects of forecast quality. *Monthly Weather Review* **124**, (to appear).

Murphy, A. H. and Daan, H. (1985). Forecast evaluation. *Probability, Statistics, and Decision Making in the Atmospheric Sciences* (A. H. Murphy and R. W. Katz, eds.). Boulder, CO: Westview Press, 379–437.

Murphy, A. H. and Winkler, R. L. (1984). Probability forecasting in meteorology. *J. Amer. Statist. Assoc.* **79**, 489–500.

Murphy, A. H. and Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review* **115**, 1330–1338.

Murphy, A. H. and Winkler, R. L. (1992). Diagnostic verification of probability forecasts. *Int. J. Forecasting* **7**, 435–455.

Pearl, J. (1978). An economic basis for certain methods of evaluating probabilistic forecasts. *Int. J. Man-Machine Studies* **10**, 175–183.

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

Roberts, H. V. (1965). Probabilistic prediction. *J. Amer. Statist. Assoc.* **60**, 50–62.

Sanders, F. (1963). On subjective probability forecasting. *J. Appl. Meteorology* **2**, 191–201.

Sarin, R. K. and Winkler, R. L. (1980). Performance-based incentive plans. *Manag. Sci.* **26**, 1131–1144.

Savage, L. J. (1954). *The Foundations of Statistics*. New York: Wiley.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* **66**, 783–801.

Schervish, M. J. (1989). A general method for comparing probability assessors. *Ann. Statist.* **17**, 1856–1879.

Shuford, E. H., Albert, A., and Massengill, H. E. (1966). Admissible probability measurement procedures. *Psychometrika* **31**, 125–145.

Spetzler, C. S. and Staël von Holstein, C.-A. S. (1975). Probability encoding in decision analysis. *Manag. Sci.* **22**, 340–358.

Staël von Holstein, C.-A. S. (1970). *Assessment and Evaluation of Subjective Probability Distributions*. Stockholm: ERI, Stockholm School of Economics.

Wallsten, T. S. and Budescu, D. V. (1983). Encoding subjective probabilities: A psychological and psychometric review. *Manag. Sci.* **29**, 151–173.

Wilks, D. S. (1995). *Statistical Methods in the Atmospheric Sciences*. New York: Academic Press.

Winkler, R. L. (1967a). The assessment of prior distributions in Bayesian analysis. *J. Amer. Statist. Assoc.* **62**, 776–800.

Winkler, R. L. (1967b). The quantification of judgment: Some methodological suggestions. *J. Amer. Statist. Assoc.* **62**, 1105–1120.

Winkler, R. L. (1969). Scoring rules and the evaluation of probability assessors. *J. Amer. Statist. Assoc.* **64**, 1073–1078.

Winkler, R. L. (1986). On "good probability appraisers". *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (P. Goel and A. Zellner, eds.). Amsterdam: North-Holland, 265–278.

Winkler, R. L. (1994). Evaluating probabilities: Asymmetric scoring rules. *Manag. Sci.* **40**, 1395–1405.

Winkler, R. L. and Murphy, A. H. (1968). "Good" probability assessors. *J. Appl. Meteorology* **7**, 751–758.

Winkler, R. L. and Poses, R. M. (1993). Evaluating and combining physicians' probabilities of survival in an intensive care unit. *Manag. Sci.* **39**, 1526–1543.

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

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

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

DISCUSSION

JAVIER MUÑOZ *(Presidència de la Generalitat Valenciana, Spain)*

Professor Winkler defines scoring rules as functions of probabilities and observations. I would like to ask two questions concerning this definition.

The first question has to do with the nature of the probabilities that are involved. The probabilities here cover the observations. However, the core of Bayesian statistics is probabilities on parameters: priors and posteriors. Posteriors may be more informative to the analyst than predictives, as this example illustrates. Suppose the analyst asks two experts if it will rain tomorrow. Following the Bayesian paradigm, both experts then construct models that depend on p, the probability of rain; assign priors; find posteriors and provide a predictive probability. Expert 1 concludes that his posterior about p is Beta$[100, 100]$, so his predictive is $P(\text{rain}) = .5$. The posterior of expert 2 is Beta$[1, 1]$ and his predictive is $P(\text{rain}) = .5$. If the analyst has to take a decision that only depends on the predictive distribution (say, taking his umbrella), then both distributions are equivalent, in the sense that they lead to the same action. But if his decision depends on the distribution of p (say, he has to ask another expert in order to obtain further information), then he needs the posteriors. I would like to hear more on evaluating non-observable distributions.

The second question has to do with the "interface" of random quantities. Suppose an expert is asked about a continuous random quantity. Asking for the full probability distribution can be too demanding if the expert is not trained in statistics. He may not know what a probability distribution is. We can ask for partial knowledge instead. Possibilities

include point estimate and discrete aproximations such as histograms and fractiles (West 1988). Using fractiles has several advantages. First, they are easy to understand. For instance, we could ask for a quantity for which the observation is equally likely to be above and below (a median). Second, they have arbitrary precision. The problem is that proper scoring rules for probabilities are no longer valid for fractiles (Cervera and Muñoz 1996).

JOSÉ L. CERVERA *(Instituto Nacional de Estadística, Spain)*

Let me congratulate Professor Winkler for his interesting review about the use of scoring rules to evaluate assessors, and thank Professor Bernardo for inviting me to discuss it.

The topic of scoring rules (SR) can be inserted in a broader one: that of *design of incentive mechanisms in a context of asymmetric information*, most extensively studied as an application to microeconomic relationships based on game theory. Indeed, a SR is a mechanism to provide incentive for the expert (or assessor) to elicit 'good probabilities'; this assessor possesses some private information (his/her subjective probability distribution) which is not known to the designer of the SR (asymmetry of information).

Examples of other topics that can be classified here are:

— the design by a firm of wage schemes to provide incentive for labour effort,
— the design by an insurance company of risk premiums to elicit buyers' personal utilities on money and damage,
— the contract by a political party with a consultant statistician to forecast accurately the results of an election.

Many other examples are reviewed by Fudenberg and Tirole (1991).

I will center my discussion on the question 'Which Scoring Rule?' that Professor Winkler asks in his paper. To formalize the answer to the question, let us introduce some concepts of design of mechanisms.

Scoring Rules as a Case of Mechanism Design. Design of mechanisms is a two-stage game, in which a first 'move' (the player who moves first is usually called 'a Stackelberg Leader') is played by an uninformed 'principal agent' P, and the second is played by one or more agents $E_1, \ldots, E_J$.

The first move consists on the design by P of a mechanism (μ, t) where $\mu = (\mu_1, \ldots, \mu_J)$ represents messages sent by $E_1, \ldots, E_J$ and $t = (t_1(\mu), \ldots, t_J(\mu))$ is a vector of transfers from P to the agents that depend on the messages sent and, possibly, on the realization of some random event. The transfers can be monetary rewards, scores, or other items.

In the second move, the agents choose $\mu_j^*(\theta_j)$, their optimal messages depending on their private information θ_j, that maximize their expected transfer. Private information θ_j can be considered as Harsanyi's type of agent (Harsanyi, 1967) .

The backward solution to the game is that the principal agent P is supposed to design the incentive mechanism optimally, in the sense that it induces the agents to send their 'true' information (or 'true feelings' in words of de Finetti), and it maximizes the *ex-ante* utility of P.

In a context of decision-making by P, the use of tranfer functions that encourage the agents not to deviate from sending their 'true' information will eliminate one source of uncertainty, as Hirsleifer and Riley (1992) point out.

In the case we are interested in, the agents are the experts, the messages are the assessed probabilities of an event, or a probability distribution on the outcomes of a random variable, and the transfer function is the SR. For example,

$$t(\mu) = -\mu^2 \quad \text{if } A, \qquad -(1-\mu)^2 \quad \text{if not } A,$$

is the quadratic SR when the message μ is the assessed probability of A by the expert. It is known that , if the utility function of E is linear in the score, he/she will choose $\mu^\star(\theta) = \theta$, where the private information θ is his/her subjective probability of A.

Note that the behaviour of P is not formally taken into account in the statistical literature about SR. Indeed, the choice of the SR is given, as Professor Winkler extensively shows, by a collection of characteristics of SRs (local SR, sensitivity to distance, difficulty adjusted SR, etc.) that guarantee the 'quality' of the assessed probability. The microeconomic literature considers P as one of the players; therefore he/she has an utility function on the actions of the remaining players. Of course, the frame of game theory is useful only if there is a conflict between the assessor and his/her client (P) in terms of rewards and losses.

The existence of several agents (experts) in the game requires the rules to combine their messages. Some examples of this topic, which I will not consider here, are the following:

— when messages are probabilities, combination can be conducted by weighting and pooling (see for example Kadane, 1993),
— when messages are 'willingnesses to pay', they may be aggregated by Clarke or Groves' mechanisms (see, for example, Fudenberg and Tirole, 1991).

Which Scoring Rule?. The above question can be now answered in terms of the solution of the problem of the design by P of a transfer function. To simplify, the set of admissible solutions is the set of piecewise smooth funtions (or functionals) of the messages, so that the problem of designing the mechanism is an optimization (or variational) problem.

The problem will be better illustrated with a couple of examples.

Example 1. Suppose that the transfer is monetary, from P to the expert, equal to the score that the expert gets after the observation of the outcome of a dichotomous $(A, \bar{A})$ event, for which he/she had assessed a probability equal to μ. Let us suppose that both P and the expert have utilities on money that are linear. The mechanism that P will choose is the solution of the problem:

$$\min \quad \int_0^1 (\mu t_1(\mu) + (1-\mu)t_2(\mu))f(\mu)d\mu$$

subject to

$$pt_1(p) + (1-p)t_2(p) \geq pt_1(\mu) + (1-p)t_2(\mu), \quad \forall p, \mu.$$

We are supposing that P will take the expert's assessed probability of A as his own. Note that $f(\mu)$ represents, as in Winkler's paper, the analyst's distribution of the assessor's p. Of course, the constraint implies that the scoring rule must be proper (in microeconomic literature, this constraint is usually called an 'incentive compatibility constraint').

Example 2. The expert is asked for a density function on the outcomes of a real random variable X, for which he reports $\mu(x)$. The scoring rule is a functional $t(\mu, x)$ that depends on the observed value x

of X and the assessor's report, $\mu(x)$. The constraint of the variational problem that denotes propriety is an integral inequalty:

$$E_p[t(p,x)] \geq E_p[t(\mu,x)], \forall p, \mu.$$

The variational problem is defined by an objective function that is the expectation of a utility function $U(S(\mu,x))$ which can be related, for example, to the decomposition

$$E_p[S(\mu,x)] = E_p[t(p,x)] + C(t,\mu,p,x)$$

where, as in Winkler's paper, $C(t,\mu,p,x)$ is a penalty for any deviation from p (calibration penalty).

The constraints are differential equations or inequalties. For example, propriety constraints for the problem of assessing the probability of a random event A can be given by the Savage representation of $t(\mu)$: t is proper if

$$\frac{t_1'(\mu)}{t_2'(\mu)} = -\frac{1-\mu}{\mu}.$$

For a multiple choice event, $(t(\mu) = t_j(\mu_j)$ if $A_j)$, t is proper if

$$\sum_j \mu_j \left(\frac{\partial t_j}{\partial \mu_k} - \frac{\partial t_j}{\partial \mu_l} \right) = 0, \forall k, l.$$

The formalization of the choice of the SR must then include the description, by means of a utility function for P, of the relative weight that P assigns to concepts such as localness, sensitivity to distance, monetary cost, or even to the predicted outcome, as in the case of the performance-based incentive plans of Sarin and Winkler (1980).

A Design Example. We will develop an example of mechanism design of a scoring rule. Suppose that P is interested in adjusting for difficulty a monetary scoring rule for the assessment of the probability of A. Utilities on money are supposed linear both for P and E. Of course, P is interested in a proper SR, so that *the second move* of an expert E with private information $Pr(A) = \theta$ will be $\mu^\star(\theta) = \theta$. Propriety can be written as

$$t_1(\mu) = g(\mu) + (1-\mu)g'(\mu), \quad t_2(\mu) = g(\mu) - \mu g'(\mu),$$

for any convex function g.

Let the prior information of P about θ be given by $f(\theta)$. Since the expected monetary transfer is

$$\theta t_1(\theta) + (1-\theta)t_2(\theta) = g(\theta),$$

the first move, that of P, will be

$$\text{minimize} \quad \int_0^1 g(\theta)f(\theta)d\theta.$$

When adjusting for difficulty, P wants $\mu = \theta_0$ to minimize the expected transfer $\theta t_1(\mu) + (1-\theta)t_2(\mu) = g(\mu) + (\theta - \mu)g'(\mu)$, for all θ, so that the first derivative with respect to μ at θ_0 will be zero: $g''(\theta_0) = 0$. The design problem is now

$$\min_g \int_0^1 g(\theta)f(\theta)d\theta$$

subject to

$$g \quad \text{convex} \quad (g''(\theta) \leq 0, \quad \forall\theta), \quad \text{and } g''(\theta_0) = 0,$$

which is a variational problem with differential inequalties and fixed end points (a Lagrange problem), and can be solved with classical methods (see for example Hadley and Kemp, 1971).

Conclusions. We have seen, by some examples, that the thought-provoking qualitative collection of characteristics of SRs reviewed by Prof. Winkler can be used, by the formalization of design of incentives mechanisms, to solve the problem of which SR to choose.

JOSÉ M. BERNARDO *(Universitat de València, Spain)*

Professor Winkler's paper is mostly concerned with the use of scoring rules in the evaluation of probabilities and does provide useful signposts to a large part of the huge related literature. An important related paper which covers some of the same ground is Dawid (1986). In this brief comment, I would like to extend those signposts, by drawing attention to the rather strong connections of scoring rules in general, and the logarithmic scoring rule in particular, with some other important topics:

Foundations. Scoring rules may be used to *define* (subjective) probability. In terms of the quadratic score, de Finetti (1963, 1964) defines the probability that an individual assigns to an event E as the number $p = P(E)$ he or she would select if he or she is to suffer a loss of $(1-p)^2$ utility units if E occurs, and a loss of p^2 utility units if E does not occur. He then shows that if a collection of such evaluations are made, those have to obey the laws of (finitely additive) probability. Important subsequent generalisations of this idea were made by Savage (1971) and Lindley (1982). A recent related reference is Eaton (1992).

Scientific inference. Bernardo (1979) suggested that Bayesian inference on a random quantity x given data $\boldsymbol{z}$ may be described as a decision problem where the action space is the class of probability distributions $p_x(. \mid \boldsymbol{z})$ which could possibly be *reported* as the final conclusion about x given $\boldsymbol{z}$, and the utility function is a scoring rule $u\{p_x(. \mid \boldsymbol{z}), x\}$ which measures the desirability of reporting $p_x(. \mid \boldsymbol{z})$ if x obtains. Scientific inference may then be characterized by preferences where the scientist is motivated to tell the truth, and where only the probability attached to the true value of x matters, i.e., by a proper, local scoring rule; under regularity conditions, this must be of the logarithmic form

$$u\{p_x(. \mid \boldsymbol{z}), x\} = A \log p_x(x \mid \boldsymbol{z}) + B(x), \quad A > 0.$$

Design of experiments. To choose an experiment is a two stage sequential decision problem, where one selects the experiment e, a result is observed $\boldsymbol{z}_e$, an action is then taken, and a terminal consequence occurs. In scientific inference, the second stage may be modelled by the structure described above. It the follows that the best experiment is that which maximizes the expected Shannon's information

$$I(e) = \int p(\boldsymbol{z}_e) \int p(x \mid \boldsymbol{z}_e) \frac{p(x \mid \boldsymbol{z}_e)}{p(x)} \, dx d\boldsymbol{z}_e.$$

It follows that Lindley (1956) criterion is a particular case of the general principle of maximizing expected utility.

Approximations. In a problem of scientific inference, if the uncertainty about a random quantity x is described by $p_x(.)$, then the expected loss of using instead an approximation $\hat{p}_x(.)$ is given by

$$\int p_x(x) \log \frac{p_x(x)}{\hat{p}_x(x)}\, dx,$$

the logarithmic divergence. It follows that this is often a most appropriate measure of approximation error in many statistical problems. For some of the implications, see Bernardo (1987).

Convergence. It follows from the argument above that an interesting *definition* of convergence of a sequence of probability distributions $\{p_n\}$, described by their densities with respect to some dominating measure μ, is often

$$\lim p_n = p \iff \lim_{n\to\infty} \int p \log \frac{p}{p_n}\, d\mu = 0.$$

Probably, some interesting research is to be done on the interrelations between this and other, more often used, definitions of convergence of probability of distributions. Is the author aware of any relevant results?

Model evaluation. Scoring rules lend themselves naturally to evaluate predictive distributions by some form of cross-validation. For instance, a natural method to evaluate the relative advantages of model M_j for prediction, given data $\boldsymbol{z} = \{x_1, \ldots, x_n\}$ is its approximated expected logarithmic score given by

$$\sum_{i=1}^{n} \log p_j(x_i \mid \boldsymbol{z} - \{x_i\}),$$

where $p_j(. \mid \boldsymbol{z})$ is the (posterior) predictive distribution which corresponds to model M_j. For some details, see Bernardo and Smith (1994, Ch. 6).

The following contributions were later received in writting.

GAIL BLATTENBERGER *(University of Utah, USA)*

First I would like to thank Robert Winkler for a succint and well written review of scoring rules and the evaluation of probabilities. As he states " the potential for greater use of probabilities and scoring rules is almost limitless" (p. 23). It is this objective I would like to address.

Both the ex ante and ex post perspectives are introduced in this article, but the ex ante perspective is emphasized. This is appropriate for a summary, but it avoids some the difficulties encountered in applications. Increased use requires familiarity with the application of the proposed tools in real world situations. Applications are generally, although not necessarily, ex post. Since expansion of scoring rule use necessitates applications, I will focus on an ex post perspective.

An assessment of the predictive performance of econometric models is a fundamental issue with which econometricians must deal. This is true not just for Bayesians, but for all econometricians. This is therefore an arena in which scoring has a potential for increased use. Assessment of predictive performance is paramount for followers of Ramsey, deFinetti, and Savage. I take a subjective predictivist position. An issue then becomes the appropriateness of scoring for the assessment of predictive performance and standards of evaluation for this process.

Problems of adjusting for differing forecasting situations are resolved when comparisons are made among model-generated forecasts for the same series in the same situations, but difficulties in forecast comparison remain. The ex ante purpose of scoring is elicitation. The ex post purpose of scoring forecasters' subjective opinions might be to encourage them to improve their forecasts. The ex post purpose of scoring model-generated forecasts is to record their forecasting performance. The predictive distributions generated from the model are themselves coherent. This is the basic normative requirement of a subjectivist model. At least from the subjectivist perspective, these coherent distributions are not right or wrong. A poor scoring model might lead an advocate of that position to feel unlucky or it might lead him to reconsider his position. There is no formal way the analyst should learn from the scoring behavior of the model. The scoring exercise cannot in this sense be evaluative. Nonetheless it is an important exercise to examine forecast performance. Good scoring models will attract followers, and scoring patterns will raise questions which will induce further research.

A record of a model's scoring performance may be presented in graphical form or in summary statistics. Graphical displays are a useful tool in this situation. A comparison of time patterns of scoring can indicate differing model performance in different periods (Blattenberger and Lad, 1988). A poor cumulative score may result from a poor performance in only one period and a superlative performance elsewhere. Summary statistics are woefully inadequate in identifying this situation. Winkler has made important contributions in the area of developing useful graphical displays (Murphy and Winkler, 1992).

Scoring rules for continuous random variables are another contribution of Winkler (See Matheson and Winkler, 1976). These rules are scores for the entire distribution, not simply its location. Econometric models commonly focus on continuous random variables and provide a forum to explore the use of these rules (Blattenberger, 1996, Blattenberger and Lad, 1988). These scores address the question that if two forecasters specify the same location for their forecasts and it turns out to be far off, should the forecaster with a high precision to his forecast receive a lower score. Scoring performance under these rules can differ substantially. The choice of scoring rule is dependent upon the loss function of the investigator, but these connections are not completely clear. Work remains to be done on the appropriateness each of these rules to particular situations and how the different rules might be benchmarked. Theoretically the choice of the rule may be tied to the investigator's loss function, but practically it seems tied to convenience and custom. Increased usage of a variety of rules would increase familiarity with their properties. The tie between the logarithmic rule and the likelihood principle is an important feature of that rule. Personally, this is an attractive feature, but I remain unconvinced of the superiority of this rule. I think Winkler is also. One feature of the scoring rules for continuous random variables which needs development is their adaptability to multivariate applications. While computationally tedious requiring numerical integration, this would be feasible based on a multivariate density function. Another issue concerning scoring rules of continuous random variables which needs investigation is that all of these rules remain symmetric.

Skill or difficulty in forecasting is another area in which Winkler has made important contributions (Winkler, 1994). This is represented with asymmetry of the scoring rule. It is simplist to illustrate properties of

scoring rules in a single event situation. Thus, the problem of comparing forecaster of the probability of precipitation in Phoenix and Portland is clear. The forecaster in Phoenix will naturally get a better score and there is a need to make an adjustment for the forecasting situation. I am wondering about the same two forecasters forecasting temperature. A squared error scoring rule extending the Brier score to a situation of continuous random variables might be used. My priors say that again the forecaster in Phoenix would look better than the one in Portland because I think there would be a more regular pattern of temperatures in Phoenix. How could this scoring rule be adjusted to reflect the difficulty of this forecasting situation? How could the skill adjustment be extended to the other scoring rules introduced for continuous random variables?

Decomposition of scoring rules is also applicable to ex post forecasting. Undue isolated attention has been focused on the calibration component to the exclusion of sharpness or refinement . An exercise in decomposing ex post forecasting distributions into calibration and refinement components for forecast deciles was undertaken in Blattenberger and Lad(1988). In this case the Brier score averaged over the deciles was fixed but its composition and the pattern of this decomposition over deciles varied substantially. This pattern was informative. More work in the application of scoring decomposition to ex post forecasting is warranted.

Returning to the question of increased use of probabilities and scoring rules, it should be clear from the above discussion that I am a user of scoring rules, in particular for the comparison of model-generated forecasts. While I strenuously avoid the use of parameter estimation and hypothesis testing in my own work, I see them and have to confess I notice them in other's work and am aware of their standards of acceptance. When I compute scoring rules in my own work I have to study them carefully to understand what is going on. Referees need to be convinced of the usefulness of graphical displays. This leads me to look for some standard methods in demonstrating what is a good score or scoring pattern. Traditional methods have thrived on the establishment of evaluation standards. On the other hand this could be a drawback. There is increased attention (McCloskey and Ziliak, 1996) to the issue of confusing statistical significance and economic importance. The application of scoring procedures to model-based forecasts forces one to

think about the questions asked and the importance of the evidence obtained. How much is increased use tied to the establishment of accepted standards and is this a direction to follow?

JOSEPH B. KADANE *(Carnegie Mellon University, USA)*

Bob Winkler has given us a fascinating review of the literature of scoring rules. I want to go over the basics briefly, because they seem to lead me to a slightly different emphasis than is apparent in Bob's review.

I take the ex ante view as fundamental, and understand the ex post as an attempt to enforce a scoring rule, ex ante, on the forecaster. In this context, the scoring rule idea seems to me to amount to an assumption that the forecaster is a utility-maximizing agent whose utility function is known. In empirical situations this assumption seems to require some caution, since it can amount to assuming that someone's motives are known without inquiry into them.

Now suppose that a forecaster's utility is known, and happens not to be proper. Need this knowledge cause us to discard this forecaster's assessments? I argue not.

To take a simple example, suppose that an assessor has a weighted quadratic utility:

$$S_1(r) = -(1-r)^2 \quad S_2(r) = -wr^2, \tag{A1}$$

for some known positive w.

Simple differentiation yields

$$p = wr/[1 - r + wr], \tag{A2}$$

which simplifies to $p = r$ when $w = 1$ (as it must, since $w = 1$ is the quadratic utility (Winkler's (3), known to be proper). In this case, knowing the assessor's announced r, his p (subjective probability) can be recovered. Thus what seems to be important here is that the mapping between r and p be one-to-one, and not necessarily that it be the identity function.

Does such a probability assessor deserve to be called dishonest? I think not. Would the knowledge that, for some forecaster, w is known and is not equal to 1 lead to some analytic difficulty? As long as (A2) is used to map back from the elicited r to the subjective p, I do not see why this causes any difficulty.

My question, then, is why is so much stress in the literature reviewed here put on propriety of scoring rules?

DENNIS V. LINDLEY *(Minehead, UK)*

In commenting on this admirable paper, I offer some critical remarks about scoring rules; not because I think the rules are unsound, quite the contrary, but because they, and the problem they address, are so important that it is essential to be clear on their foundations.

Winkler mentions one difficulty, that all rules implicitly assume the assessor's utility function is linear in the score. As he remarks, if the function is unknown, the assessor's statement loses much of its value. The difficulty goes deep. Probabilities can only be assessed in the context of a decision problem. (The assessor's choice of r is a decision about p.) For Bayesians, who hold that decision analysis for a single decision-maker consists in maximizing expected utility, MEU, this means that the utility function $u(d, \theta)$, for decision d and uncertain θ, and probability $p(\theta)$, always appear as a product. That is, probability can never be separated from utility, at least not without assumptions additional to those ordinarily used to justify MEU. Schervish (1995) introduces one in his axiom 4 of state independence (p.184). One way to assess your probability for an event E is to consider an urn with a proportion p of white balls, and allow you to select p such that the uncertainty of E is the same as the uncertainty of a white ball when drawing a ball at random. But this is only easy to use with events you consider ethically neutral. Replacement of the uncertainty of nuclear war with withdrawal of a ball is not innocuous. Although I think belief has a meaning for me separate from actions based on that belief, I find it hard to devise a reliable form of measurement for that belief that is separate from utility considerations. Rubin (1987) is relevant.

Suppose the utility function is known to be linear and the scoring rule thereby becomes effective, why should it be proper? I showed, Lindley (1982), that, aside from some pathological cases, *every* scoring rule encourages the assessor to state a monotone function of probability; a function which depends only on the rule. For example, the rule $S_1(r) = e^{-r/2}, S_2(r) = e^{+r/2}$ will yield r as log-odds. Expressed differently, every scoring rule is proper for something. Log-odds is just as valuable as probability in expressing uncertainty. The proof uses the weaker

concept of admissibility in place of expectation. This does not avoid the brush with utility since the proof is based on the addition of scores.

The role of the likelihood principle in the context of scoring rules is unclear to me. One form of the principle, Berger and Wolpert (1984), says that if $E = (X, \theta, \{f_\theta\})$ is an experiment, then $Ev(E, x)$ should depend on E and x only through $\ell_x(\theta)$. Here X is the quantity to be observed and x its observed value, θ the parameter indexing the family $\{f_\theta\}$ of possible densities for X, Ev the evidence about θ provided by E when it yields x, and $\ell_x(\theta)$ is the likelihood at $X = x$, that is, the set of densities at x as θ ranges throughout its whole range. In the context of scoring rules, E consists in obtaining the assessor's judgement r and then observing the true state, t say, so $x = (r, t)$. θ is the assessor's probability p. For the analyst, the likelihood is presumably $pr(r, t|p)$ as a function of p, where pr denotes the analyst's probability. Winkler suggests that the principle implies only local rules but I do not see how this follows from the principle as just described. If $r = (r_1, r_2, \ldots r_n)$ expresses a view about a partition $A = (A_1, A_2, \ldots, A_n)$ and $t = A_j$, why should $pr(r, A_j|p)$ reduce to $pr(r_j, A_j|p)$, or the expected score to a function only of r_j and t?

Perhaps Winkler has a different likelihood in mind. That just given fits into a framework in which an analyst is using the assessor's judgement r. The analyst is then interested in $pr(A|r) \propto pr(r|A)pr(A)$. In the case of a dichotomy ($n = 2$) the likelihood involves $pr(r|A = 1)$ and $pr(r|A = 0)$ and only their ratio matters. In frequency calibration, attention focuses on $A = 1$, given r, which seems to have the quantities in the wrong order. In this approach, it is not true that a well-calibrated probability, which is the same as the analyst's, will leave the latter unaltered, as is claimed in Section 6.

The Bayesian view of the world is one in which the emphasis is on coherence. It prescribes how the different uncertainties you have must fit together, according to the convexity, addition and multiplication rules. But it is silent on what those uncertainties should be. It is like geometry, that says the angles of a triangle must add to 180 degrees, but says nothing about what these angles are. To determine them, there is a technology of mensuration, based on theodolites and other apparatus. We lack the theodolites of probability, where scoring rules provide a possible answer. Nevertheless, I feel that coherence itself must play an important role in

our measurement. The empirical literature, for example in meteorology, is concerned with repetitions of a single event, like 'rain tomorrow'. But what experience is there of asking assessors, who understand coherence, for the probabilities of several, related events? For example, with just two, A and B, what about probabilities for the partition, AB, AB^c, A^cB, A^cB^c, and for sequences A, B given A, B given A^c, and then with the roles reversed? All these must cohere and surely that coherence should be invoked, despite its failure to relate, as a scoring rule does, to the actuality of A and B. Furthermore, if, for example, the quadratic rule is used, it will provide different scores if the partition is requested from those when a sequence is required. Is one better than the other?

Scoring rules offer a most promising way of relating coherence to reality. The present paper advances our understanding of this relationship significantly and it is no criticism of it to say that many questions remain unaswered.

ALLAN H. MURPHY *(Prediction and Evaluation Systems, USA)*

Introduction. It is a pleasure to have an opportunity to comment on Robert Winkler's excellent review (Winkler, 1996) of scoring rules (SRs) and related measures for evaluating probabilities. My comments will focus on some of the issues raised in his paper from the viewpoint of an individual who is primarily concerned with the evaluation of probability forecasts (PFs) in a meteorological setting. Nevertheless, the comments touch upon issues that should be of at least some interest to individuals concerned with evaluating the goodness of PFs in all settings, whether these forecasts are based on subjective judgments or on so-called "objective" models (or a combination of judgments and models).

Probability Forecasting and Probability Forecasts in Meteorology. It may be useful to start by describing briefly the current status of probability forecasting in meteorology. Today, PFs of future weather conditions range from relatively detailed, event-specific predictions for the next few hours to forecasts of mean temperature or median precipitation anomalies for periods of a month or a season up to a year in advance. Four approaches to probability forecasting can be identified (Murphy and Ehrendorfer, 1996): (1) subjective probability forecasting; (2) statistical probability forecasting; (3) numerical-statistical probability forecasting;

and (4) stochastic-dynamic probability forecasting. These methods are frequently used in various combinations to produce the "official" PFs (and non-PFs) disseminated to users of weather forecasts.

To clarify, subjective PFs are produced by forecasters, usually with the aid of numerical and numerical-statistical model output as well as observed data. Statistical PFs, generally limited today to very short and very long lead times, are derived solely from statistical models (e.g., regression models, time-series models). Numerical-statistical PFs are produced by the joint use of so-called "numerical" (i.e., physical-dynamical) and statistical models; specifically, the output of numerical models is used as input to (i.e., as predictors in) the statistical models. Finally, stochastic-dynamic PFs are now produced on a quasi-operational basis under the rubric of ensemble forecasting. In this approach, a numerical model is integrated forward in time from an ensemble of — for example, 25-50 — carefully selected initial conditions (under the assumption that the "true" initial conditions are imperfectly specified), yielding an ensemble of future weather conditions from which empirical PFs can be derived.

It may be of interest for readers to know that meteorological forecasting systems (employing the above-mentioned methods) now produce millions of PFs each day on a worldwide basis. For example, the numerical-statistical forecasting system in the United States now formulates hundreds of thousands of PFs daily (various combinations of weather variables, geographical locations, lead times, etc.), and national weather services in several other countries (e.g., Canada, France, Sweden) also produce many such forecasts on a routine basis. Not all PFs are available for "public" consumption; they are frequently intended primarily as guidance to forecasters. Nevertheless, the recent increase in the commercial activities of both public and private weather services has resulted in a greater interest in PFs, with the result that national weather services in some countries (e.g., Sweden) now provide most of their users with forecasts in a probabilistic format. Unfortunately, a lamentable tendency still exists in some circles to translate PFs into non-PFs before they are disseminated to the user community.

Scoring Rules in Probability Assessment. As noted by Winkler, strictly proper SRs provide an incentive for forecasters to report their probabilities honestly. To some degree, this feature of the Brier score (BS) (Brier, 1950) accounts for its popularity and widespread use as a summary measure of the accuracy of PFs in meteorology. However, in the operational weather forecasting milieu with its severe time constraints, the incentive role of the BS has often taken a "back seat" to its role as an ex post evaluation measure. In effect, weather forecasters formulate their PFs intuitively without the aid of formal probability assessment methods, but with strong input from probabilistic and nonprobabilistic guidance.

It should be noted that the role of any strictly proper SR in a meteorological context is clouded by the fact that forecasters are frequently influenced by their perceptions regarding the impacts of the weather itself and/or the effects of weather forecasts on users, particularly in situations involving hazardous weather. In such situations, the forecaster's utility function may no longer be linear in the SR (e.g., the BS) and — to encourage honest reporting — the BS should be replaced by a SR that takes this nonlinearity into account. Winkler's discussion in Section 6 elucidates some of the issues involved in these situations.

Evaluation of Probability Forecasts. Perhaps it is useful to expound briefly on the perspective of the writer as regards the evaluation of PFs (and all forecasts for that matter). This perspective derives primarily from a paper by Murphy and Winkler (1987) (hereafter MW87), as well as the writer's subsequent efforts to elaborate on a framework for forecast evaluation originally set forth in MW87 (e.g., Murphy, 1991, 1995). Under the assumption that the bivariate time series of forecasts and observations (for a single forecasting method or forecaster) is uncorrelated and stationary, the joint distribution of forecasts (r) and observations (x), $p(r, x)$, contains all of the information relevant to the quality of the PFs (thus, quality represents the totality of the statistics regarding the forecasts, the observations, and their relationship that can be computed from this joint distribution). For the purposes of these comments, it is assumed that evaluation issues related to temporal dependence and nonstationarity are addressed separately.

Two further points are worth mentioning here. First, as noted in MW87, the joint distribution $p(r, x)$ can be decomposed in two differ-

ent ways into conditional and marginal distributions; namely, (1) $s(x|r)$ and $u(r)$ and (2) $t(r|x)$ and $v(x)$. These latter distributions contain important information in an evaluation sense, since they can be related to specific attributes (or aspects) of forecast quality. Second, consideration of these distributions and their fundamental role in the process of evaluating PFs leads to the identification of a basic characteristic of evaluation problems referred to as *dimensionality* (Murphy, 1991). Under the assumption that the PFs of interest are to be evaluated on the basis of the empirical joint relative frequencies of forecasts and observations (for simplicity the estimation problem is ignored here), the dimensionality of an evaluation problem is simply one less that the product of the number of distinct forecasts times the number of distinct observations. Thus, an evaluation problem involving 11 distinct PFs (or ranges of probability values) for a binary variable is a 21-dimensional problem. It takes 21 parameters (i.e., relative frequencies) to reconstruct $p(r, x)$. [Note: Instead of using empirical relative frequencies, these joint distributions could be modeled with parametric statistical models (e.g., Murphy and Wilks, 1996). This approach reduces the dimensionality of the evaluation problem and may also "smooth out" some of the effects of sampling variability. In meteorology, such models have been used heretofore primarily in decision-analytic studies concerned with the value of weather forecasts (e.g., Katz, Murphy and Winkler, 1982).] When PFs are evaluated using a summary (i.e., scalar or one-dimensional) SR, many aspects of the quality of the forecasts are necessarily confounded in this measure of performance.

Evaluation Measures and Scoring Rules. It may be of some interest to emphasize the point (implicit in Winkler's Section 1) that the class of scoring rules is a subset of the class of all potentially relevant evaluation measures. SRs are defined such that a meaningful score can be calculated for an individual PF and the corresponding observation (e.g., Murphy, 1996). Thus, measures of aspects of quality defined in terms of averages over samples or subsamples of forecasts and/or observations do not qualify as scoring rules. Examples of the latter include the usual measure of unconditional or systematic bias (e.g., the difference between the mean forecast and the mean observation), as well as the measures of reliability (calibration) and resolution in a decomposition of the ex post Brier or quadratic score (see Section 6 in these comments). From this

perspective alone, it may be unwise to limit the evaluation of PFs to the calculation of a summary SR.

Decompositions of Scoring Rules. In the opinion of the writer, decomposition of SRs (see Section 4 in Winkler's paper) is important for at least two interrelated reasons. First, decompositions provide quantitative measures of aspects of quality that are confounded when evaluation of forecasting performance is limited to the overall SR itself. Second, evaluation of probabilities based on a SR and the components of a decomposition of this SR at least implicitly recognizes that evaluation problems are multidimensional in nature and that forecast quality consists of several distinct attributes.

Consideration of one of several possible decompositions of the average ex post BS, E(BS), in the case of PFs for a binary variable ($x = 0$ or 1) may be illustrative here (Murphy, 1973; Murphy and Winkler, 1992):

$$E(BS) = E_x[x-E(x)]^2+E_r[r-E(x|r)]^2-E_r[E(x|r)-E(x)]^2. \quad (1)$$

As defined in (1), $E(BS)$ is the negative of the overall average quadratic score considered by Winkler [this version of $E(BS)$ is sometimes referred to as the average half Brier score]. The terms in (1) are, respectively, the variance of the binary observations (a characteristic of the forecasting situations rather than of the PFs), a measure of calibration [the PFs are perfectly calibrated if $E(x|r) = r$ for all r], and a measure of resolution [the PFs are completely unresolved if $E(x|r) = E(x)$ for all r]. Thus, a measure of overall accuracy $[E(BS)]$ has been decomposed into a measure of variability in the forecasting situations and measures of reliability (calibration) and resolution of the PFs. Evaluation and/or comparison of forecasting performance in terms of all of these measures is potentially more informative than evaluation in terms of E(BS) alone. Moreover, this approach is consistent with the spirit (if not the full dimensionality) of the evaluation problem. In the case of perfectly calibrated forecasts (in a frequency sense), the reliability term vanishes and the resolution term becomes a measure of refinement or sharpness {i.e., $E_r[E(x|r) - E(x)]^2 = E_r[r - E(r)]^2$}.

Comparability of Scores. In connection with his discussion of the comparability of scores (Section 5), Winkler mentions skill scores, which (as he correctly notes) are widely used in meteorology as a means of comparing forecasting performance across different situations (e.g., geographical locations, time periods). For example, a skill score based on the Brier score [expressed in a form analogous to Winkler's equation (23)] is frequently used to compare overall probability forecasting performance. Winkler's family of strictly proper asymmetric scoring rules clearly represents an interesting alternative to traditional skill scores used in meteorology, especially since comparative evaluation based on these two types of skill scores may yield quite different results (Winkler, 1994).

In judging the results of a comparative evaluation based on skill scores (or any other one-dimensional measure of forecasting performance), two considerations should be kept in mind. First, since such skill scores generally do not describe forecast quality in its full dimensionality, superiority in skill (or any other single aspect of quality) is no guarantor of sufficiency (see Section 10 of these comments). Second, skill scores are intended to "level the playing field" and thereby enhance the validity of comparisons across forecasting methods or forecasters. However, these skill scores generally take into account only differences in the climatological (or base rate) probabilities of the events in the two (or more) sets of forecasting situations. Many other factors could contribute to "forecast difficulty" at different locations or in different time periods, including (for example) the sequence of weather regimes and the day-to-day persistence in the weather (i.e., various other task characteristics). Thus, it is very unlikely that skill scores (however defined) can create a level playing field. Whenever possible, it is strongly advisable — and greatly simplifies the complexity of the evaluation problem — to compare forecasting methods or (if necessary) forecasters across the same set of forecasting situations (e.g., see Murphy, 1991).

Sensitive-to-Distance Scoring Rules. In his discussion of sensitive-to-distance SRs (Section 6), Winkler mentions the ranked probability score (RPS) formulated by Epstein (1969). This strictly proper SR is used (instead of the BS) by meteorologists to evaluate PFs in situations involving variables defined in terms of three or more categories, whose categories (or values) possess an underlying natural order (i.e., in which

the concept of distance is meaningful). It is easy to show that the average RPS is (linearly related to) the sum of the squared differences between the cumulative distribution of the PFs and the cumulative distribution of observations, whereas the average BS is the sum of squared differences between the (noncumulative) forecast and observed distributions (Murphy, 1970). In addition to the references cited by Winkler, Staël von Holstein and Murphy (1978) describe a family of strictly proper, sensitive-to-distance, quadratic SRs that can be tailored to particular applications.

Feedback and Learning. In his discussion of feedback and learning in Section 6, Winkler quite properly — in the view of the writer — emphasizes the benefits of decompositions of SRs and the use of a diagnostic approach to the evaluation of PFs. Scoring-rule feedback by itself would seem to be of relatively limited value in situations in which attention is focused primarily on the removal of biases (conditional or unconditional) in PFs or on the improvement of various desirable aspects of forecast quality (e.g., Yates, 1994). In this regard, feedback consisting of a model's or a forecaster's empirical calibration function [i.e., $s(x = 1|r)$ versus r for all r] would appear to provide the kind of information that could enable modelers/forecasters to improve the reliability of their PFs in the future (e.g., see Murphy and Daan, 1984). To give a modeler or forecaster a complete picture of forecasting performance, these calibration functions should be accompanied by a refinement or sharpness diagram [i.e., a plot of u(r) versus r]. Such diagrams, as well as discrimination or likelihood diagrams [i.e., $t(r|x = 0)$ and $t(r|x = 1)$ versus r], together with measures of these aspects of quality derived from decompositions of (quadratic) SRs, should be helpful in identifying basic strengths and weaknesses in forecasting performance. Moreover, this type of information would undoubtedly be of even greater benefit to modelers/forecasters if — by augmenting the evaluation data base — the results could be stratified by relevant covariates (an example of such a covariate in a meteorological context would be the weather regime that prevailed when the PF was formulated; see Murphy, 1995).

Superiority and Sufficiency. The relationship between superiority, as determined by a particular strictly proper SR, and sufficiency (or unambiguous superiority), as determined by the sufficiency relation (DeGroot and Fienberg, 1982, 1983), raises some interesting and challenging questions in the arena of comparative evaluation of PFs. It seems clear that the fact that forecasting method A is superior to forecasting method B, in terms of an individual strictly proper SR, is generally inadequate to ensure that A's PFs are sufficient for B's PFs. [Note: It is only under relatively restrictive conditions that an evaluation measure can be shown to be consistent with the sufficiency relation (Krzysztofowicz, 1992).] In fact, we speculate here that the proper inference in this case is that B is not sufficient for A. However, is this latter inference appropriate for the class of all (strictly proper, proper, and improper) scoring rules as well as for the even larger class of evaluation measures? Conversely, does the fact that method A's PFs are sufficient for method B's PFs guarantee that all scoring rules and evaluation measures involving different aspects of quality will indicate (ex post) that A's forecasts are superior to B's forecasts?

Probability Forecasts, Scoring Rules and Decision Making. When considering issues related to PFs, SRs, and decision making in a meteorological context, it is necessary to understand the rather special circumstances that arise in this arena with respect to PFs and their use by decision makers or endusers in the public and private sectors. First, a meteorological PF may be used by several individuals, and in many if not most cases these endusers are not idealized or sophisticated decision makers in the prescriptive decision-analytic sense. In particular, they generally do not have at hand the information needed to calibrate unreliable PFs nor can it be assumed that they would be able perform this task even if this information were available. It is for this reason that well-calibrated PFs are of considerable importance, since reliable PFs allow endusers to act on the basis of the forecasts themselves without the need for recalibration. From the perspective of this very practical setting, approaches to the evaluation of PFs, including decompositions of scoring rules, that ignore calibration — or assume that all PFs are well-calibrated — have relatively limited appeal.

Conclusion. In conclusion, strictly proper SRs can be expected to continue to play important roles in probability assessment and evaluation. In the arena of probability evaluation, the use of a SR as an overall measure of the quality of PFs is consistent with the common (but somewhat limited) objective of summarizing forecasting performance in terms of a single number. However, this writer believes that individual SRs are seriously deficient when the objective of evaluation is broadened to include the diagnosis of forecasting strengths/weaknesses and the improvement in various aspects of the quality of PFs. Moreover, comparative evaluation of forecasting methods or forecasters based solely on strictly proper SRs can be misleading since superiority in terms of overall scores is no guarantor either of unambiguous superiority across all aspects of quality or of sufficiency. Many issues of a methodological and practical nature remain to be resolved with regard to SRs and other evaluation methods/measures, and Robert Winkler's review paper provides valuable insight into recent developments and current thinking in this arena as well as into possible future directions in scoring-rule research and applications.

ROBERT M. OLIVER *(University of California at Berkeley, USA)*

I have only two comments after reading a most enjoyable paper. The second one is a question.

First comment. I wish that Winkler could add a footnote to this paper, and future Bayesians would distinguish carefully in their publications, between scoring rules (the subject of this paper) and the widespread use of scores in the financial community to predict the risk of default or bankruptcy or credit performance. The latter denote the log of posterior odds of the appropriate performance conditional on the score. As far as I can surmise there is little if any connection between the two.

My second comment and question: Suppose that a strictly proper and coherent score S is developed for a mutually exclusive and exhaustive set of events, A (a vector). Furthermore, the scoring rule is blessed by all attendees at the April 23rd 1996 meeting! An honest assessor H uses this score to test, calibrate, evaluate and validate his probabilities before he sells them, on a daily basis, to a decision maker D. D uses H's probability forecasts to make economic decisions on his business whose risk is known to be captured in the uncertainty of A. D's objective is to

maximize expected profit but he finds his profits decrease dramatically after using H's forecasts.

The moral of this story seems to be that either

(i) Strictly proper scores are not necessarily useful to decision makers, or

(ii) There is an impossibility theorem for this story.

My question is: Which is it and Why?

DAVID RÍOS-INSUA *(Universidad Politécnica de Madrid)*

Professor Winkler's paper is an excellent review on an aspect of the too frequently forgotten problem of probability elicitation and assessment. In spite of being so fundamental in both decision analysis and inference, there is still much to discover concerning ways of belief (and preference) elicitation.

My discussion will center around the elicitation aspects of scoring rules. I enjoyed the clear distinction made between the use of scoring rules for elicitation and evaluation, which has frequently been confused. It is curious that scoring rules were introduced as elicitation tools and later used as evaluation tools. Interestingly enough, textbooks and protocols tend to emphasise other elicitation methods, like those based on lotteries, with scoring rules used mainly for evaluation (see e.g. Cooke, 1991).

Having said this, a first question that comes to mind is how does scoring rule based elicitation compare with, e.g., lottery based methods. In principle, we could raise some operational/psychological doubts about the first ones. Since the betting scheme in the scoring rule method is far from the client's problem, we could induce some disinterest in him; similarly, while explaining to him that a virtue of scoring rules is to promote honesty, we could actually build some distrust in the client. More technical remarks can be made. For example, as the author describes in Kadane and Winkler (1988), scoring rule methods require stronger assumptions to ensure that the stated probability is the true probability, to account for the effects of utilities.

One intriguing thing in the paper is the role of the density f, introduced when p is not known (almost always?). It seems difficult to assess f, to say the least, much as Prof. Winkler says that the assessment of U is difficult. When f assessed by to the analyst it seems that we are giving

him too much of a role; after all, the analyst could be himself dishonest, and we might be at the beginning of an infinite regress. As the assessor is concerned, f could come from past performance in similar situations, but this may be difficult to hold in many cases. It may happen that we are only able to elicit a class of densities f; in this case, we could compare assessments with respect to that class by comparing expected scores.

Another possible use, when several experts are available, would be that each expert evaluates the others with their stated values as the relevant p's. The cutting points of their expected scores could be used as starting point to discuss their assessments as a way to look for consensus.

A number of potential areas of research and applications could be considered. For example, Winkler suggests that scoring rules may not be applied to the evaluation of probability statements about unobservables in a statistical model. Many prior elicitation methods, see e.g. Chaloner et al (1993), rely on the predictive distribution to infer the prior parameters. Hence, we could score the unobservables based on the observables. These scores would be, of course, conditional on the model.

This points out the problem of evaluating multivariate assessments based on conditionals. Conceptually, the problem is the same, but we could score the assessment conjointly, by marginals, or by conditionals. An important application would be the validation of probabilistic expert systems, Pearl (1988). In spite of the increasing interest in these systems, most validation work is still dealing with deterministic systems. Spiegelhalter *et al.* (1993) outline some uses, but there is still much work to do in this area. The final objective would be to detect when the system is not functioning correctly, and if so, which assessments require improvement.

Another potential area of interest is to extend rules to the case in which there is imprecision in the assessments, the typical situation in robust Bayesian analysis, see e.g. Berger (1994). For example, we could get bounds on the scores when the prior runs through a class of priors as a sensitivity measure. Alternatively, for ϵ-contaminated classes we could use the base measure as the "true" distribution and score the other distributions in the class. Since there is a fair deal of knowledge about the interpretation of scores, we would have a neat way of obtaining interpretable sensitivity measures, one of the pending issues in robust Bayesian analysis. That would refer to comparison of distributions as

a whole, but the solution would depend on the estimation or decision problem at hand. This puts us on the same track of some of the comments by Professor Winkler, concerning the need to relate the scoring problem to the particular decision problem.

REPLY TO THE DISCUSSION

I am grateful to the discussants for their extensive and valuable comments regarding scoring rules and related issues. They have significantly extended the signposts provided in my paper, have brought new issues and perspectives to the table, and have given us all some ideas to contemplate. Many important questions remain unanswered, as Lindley notes, which is why additional work is desirable. The discussants have identified and helped to clarify many of the most challenging questions. In that spirit, and given the many details in the discussion, I will focus on a number of key issues in my reply rather than try to acknowledge and respond to every point raised by the discussants.

Coherence. Bernardo points out that scoring rules can be used to provide an operational definition of subjective probability, and the primacy of coherence in subjective probability is worth emphasizing. Blattenberger, Lindley, and Ríos raise the issue of evaluating multivariate assessments, which are more prevalent (and more complicated) as we see more large probabilistic systems such as the expert systems mentioned by Ríos. Should we evaluate the full joint probabilities, or should we evaluate marginal and/or conditional probabilities individually?

If left to their own devices, assessors are likely to provide inconsistent probabilities, especially when we move beyond the simplest situations, and it is important to identify and remove inconsistencies as part of the elicitation procedure. Edwards and von Winterfeldt (1986, pp. 113-114) welcome such inconsistencies as an opportunity to fine-tune the assessments: "Good elicitation practice is never to rely on one way of asking. Instead, ask the same or related questions in various ways, looking for inconsistencies. If you find some, be glad. They can be fed back to the respondent, who must then be asked to think some more in order to eliminate them. Anything that promotes hard thinking and insight helps." I agree, and therefore prefer to assume that coherence will be obtained separately from any evaluation considerations rather than being motivated by evaluation considerations (e.g., by noting that inconsis-

tencies will reduce an assessor's expected score). But that still leaves the question of how to evaluate multivariate assessments, for which my quick answer is that ideally we should evaluate the joint probabilities so that the assessor's judgments about dependence will be considered in the evaluation. A slightly longer answer is that we should supplement the evaluation of multivariate probabilities with evaluations of marginal and conditional probabilities in order to understand where an assessor "does better" and where the assessor has difficulties.

Why Proper Rules? Kadane and Lindley ask why scoring rules should be proper. In theory, they need not be proper as long as the probabilities of interest can be recovered from the assessor's responses. In practice, it is hard enough to get assessors to think in terms of probabilities and to utilize their knowledge and information to come up with probabilities. To ask them to add another step of figuring out how to respond optimally to a non-proper rule can add a layer of complexity that could be bewildering to many assessors. I doubt that an assessor, who is not likely to write down and maximize expected utility formally, would come up with the optimal response, $r = p/(w - wp + p)$, in Kadane's example.

That said, response modes such as odds or log odds may be readily understandable, and the jury is out on which response mode (e.g., probabilities vs. odds vs. log odds) might lead to "better" probabilities. Lindley points out that these response modes are equally valuable, and the question is one of ease and effectiveness of use by real assessors. Of course, rules such as the linear rule, for which the assessor's probability cannot be recovered from the response (which should be zero or one), are not satisfactory.

Game-Theoretic Considerations. I am particularly grateful to Cervera for reminding me that I was remiss in not including economics in the first paragraph of Section 7 and in not mentioning the important contributions from agency theory (see also Pratt and Zeckhauser, 1985). We can learn much from the approach discussed by Cervera. I do worry about practical difficulties caused by having to solve very complicated games once we get beyond the simplest situations and by the fact that we do not generally know the assessor's utility function or other stakes the assessor might have in the events or variables of interest. Also, the formal analysis makes heroic assumptions about perfect rationality, modeling,

and computing on the part of the assessor. But to the extent that these practical difficulties can be overcome, this approach is highly relevant in formalizing the design of scoring rules for specific decision-making problems.

The primary game-theoretic consideration in agency theory involves the game between a principal and an agent (in probability assessment, a decision maker or analyst and an assessor). In situations with multiple assessors for the same set of probabilities, a game among the assessors also comes into play. As Ríos notes, each expert evaluates the others. This brings in new complications. For example, assessors may attach high utility to winding up in first place (i.e., getting the highest score). Maximizing the chance of getting the highest score is likely to motivate a strategy very different from trying to maximize expected score.

Murphy mentions yet another game-theoretic complication: the common situation in which an assessor's probabilities may be used by multiple decision makers (i.e., multiple principals) with different decision-making problems and utility functions. With public forecasts such as precipitation probabilties, the assessor not only doesn't know the utility functions or the details of the decision-making problems of the users of the forecasts; he doesn't even know who most of them are! With private forecasts provided for hire, the assessor obviously knows something about the various decision makers, but the decision makers may not know much about each other, and the presence of other principals raises the question of an assessor having other (often unknown or at best ambiguous) important stakes .

Difficulties Associated with Utility. Lindley says that the difficulty with utility goes deep, and I agree. Some problems related to utility have already been mentioned in this reply. As Kadane notes, an assessor who does not set r equal to p because of a nonlinear utility function is being perfectly rational, not dishonest. The trouble is that if we do not know the assessor's utility function, we cannot recover p. Unfortunately, we seldom know the assessor's utility function, and trying to assess it adds new difficulties (including more game-theoretic considerations). By keeping any scoring-rule-related stakes small, we are likely to be dealing with minor nonlinearities, and the assessments might be quite robust to such nonlinearities. However, this means that any non-scoring-rule-related stakes will take on even greater significance.

Priors, Posteriors, and Predictives. Muñoz states that "the core of Bayesian statistics is probabilities on parameters." This highlights the distinction between probabilities for observables and probabilities for unobservable parameters. Parameters are artifacts of models, and except for the simplest models, they are often very difficult to think about intuitively. For example, at first glance the coefficients in a multiple regression model might seem easy to think about, but when we try to assess probability distributions for them, we find that this is not a simple task. One way around this problem is to follow the spirit of de Finetti (and evidently of Blattenberger as well) by assessing predictive distributions from which we can make inferences about the underlying prior distributions (Winkler, Smith, and Kulkarni, 1978; Kadane et al., 1980). As Ríos notes, we can indirectly score the unobservables based on the observables.

This all implies that if two models (or two experts' posteriors) lead to the same predictive probabilities, as in the example given by Muñoz, we cannot distiguish between them by evaluating them on the basis of an observation of rain or no rain. In fact, we might ask what p really is in this example. It sounds suspiciously like one of those "true probabilties" which do not exist according to de Finetti. I prefer to think of it as a convenient modeling artifact and feel that although probabilities on parameters are an essential part of Bayesian parametric models, predictive probabilities are the most important and most valuable probabilities in Bayesian models because they connect these models with reality.

Probability Elicitation Methods. As noted in the above discussion of coherence, scoring rules can be used as an elicitation method. Ríos asks how they would compare with other elicitation methods, and the evidence is scanty on this point. This is a fruitful direction for experimentation. A related issue is the tailoring of scoring rules to other elicitation methods that are commonly used. The fractile approach discussed by Muñoz is a very appealing method that is widely used in probability elicitation in decision analysis. Cervera and Muñoz (1996) investigate proper scoring rules for fractiles, and another way to approach this issue is via a modification of the scoring rules developed in Matheson and Winkler (1976) for continuous distributions. Each assessed fractile can be the basis for a score based on a dichotomous situation: the variable is either above the fractile or not above the fractile. (Note that fractiles

can be assessed by giving a cumulative probability and assessing the corresponding value of the variable or by giving a value of the variable and assessing the corresponding cumulative probability.) Then we can add the scores from a dichotomous rule for all of the fractiles, even weighting them to give more attention to certain fractiles than others if desired. If we want to fit a continuous distribution to the fractiles, we can use scoring rules for continuous distributions to evaluate the fitted distributions and get bounds on the scores for classes of distributions, as suggested by Ríos.

Local Scoring Rules. The logarithmic scoring rule has a lot going for it, as Bernardo points out. (Regarding his interesting query on definitions of convergence, I am not aware of any relevant results.) One appealing characteristic is that it is consistent with standard Bayesian model evaluation via Bayes factors, which (in response to Lindley) is what I have in mind as a (perhaps imperfect) analogy with the likelihood principle. The issue is whether probabilities (assessed subjectively, determined from models, or from whatever other source) for events or values of a variable other than the event or value that actually occurs are relevant. This, to me, is the strongest argument in favor of local rules. Nonetheless, as Blattenberger perceptively notes, I remain unconvinced, as she does, of the superiority of the logarithmic rule.

Feedback and Learning/Improvement. An important ex post aspect of scoring rules is, as Kadane suggests, to enforce the ex ante motivation on the assessor. But I agree with Murphy that an even more important ex post role is to provide feedback and to encourage learning on the part of an assessor or model improvement on the part of a model builder. (In the point regarding model building, I deviate from Blattenberger.) Here, as Murphy notes, the decomposition of scores into terms involving specific attributes such as calibration and sharpness and the consideration of graphs and measures related to the joint distribution of forecasts and outcomes are much more valuable than just the overall score. Blattenberger wisely emphasizes the advantage of graphical displays over summary statistics in this respect. Extensive diagnostic analyses of this type have the greatest potential for helping us understand characteristics of probability forecasts from assessors or models (or even from combinations of probabilities from multiple assessors and multiple models) and for leading to serious improvement in the probability forecasts over time.

Of course, feedback cannot by itself guarantee expertise. Oliver lists two possible morals for his story, but there are more appealing and credible explanations. For example: the assessor, though honest, well-calibrated, and well-intentioned, may simply not be an expert on A; the "state of the art" may not permit sharp enough probability forecasts about A to enable D to make money; D may not use the assessor's forecasts appropriately (e.g., his model of the decision-making problem may be flawed); or the decision maker might simply have bad luck despite a good expert and a good process (the classic distinction between good decisions and good outcomes).

Calibration/Attributes of Probabilities. Not everyone agrees as to the relative importance of different attributes of probabilities. To my mind, Blattenberger's comment that undue isolated attention has been focused on the calibration component to the exclusion of sharpness or refinement is right on the money. Particularly in behavioral studies, little attention has been given to anything but calibration. A weather forecaster giving probability of precipitation forecasts can be well calibrated by just giving climatology as the probability forecast each day, and this approach demonstrates no real expertise. For discrimination, which is arguably the most important attribute of probabilities from a decision-making viewpoint, the actual number that is assigned as a probability is only important in that it places that probability in a particular class of probabilities assigned the same (or a similar) number. If the event of interest always occurs when the probability assessed for it is 0.2, and we are aware of this, then for us a probability of 0.2 is as valuable as a probability of one from a well-calibrated assessor. Yet Murphy's point that many users are not sophisticated enough to recognize this and to correct for miscalibration is well taken.

Also on calibration, I thank Lindley for clarifying the implications of a careless statement of mine: "... a forecast providing a well-calibrated probability that is the same as the decision maker's own prior probability is not valuable because it will not change the preferred course of action." In this situation, the decision maker's probability may indeed change after seeing the forecast, although under certain conditions it will remain the same. See Winkler (1986) and Dawid, DeGroot, and Mortera (1995) for relevant discussions.

Comparability of Scores. Unfortunately, I agree with Murphy's comment that it is very unlikely that skill scores (however defined) can create a level playing field. My hope is that they can create a playing field that, while not perfectly level, is a bit less uneven. Blattenberger asks about extensions of asymmetric scoring rules to multiple events or continuous situations. This can be accomplished by the same approach discussed above for scoring a series of fractile assessments. In the multiple event situation, use an asymmetric dichotomous rule for each possible event (for all but one event, the outcome will be "event did not occur"), then sum (or average) the scores across all events. In the continuous case, the sum is replaced by an integral, by analogy with Matheson and Winkler (1976).

Applications. Clearly my lament in Section 7 about the dearth of probabilities in real- world forecasting situations does not apply in meteorology. As Murphy points out, weather forecasting has provided many valuable data sets of probability forecasts (including subjective probabilities and model-based probabilities) and is one field where probabilities are regularly communicated to the general public and evaluated via scoring rules. Probability forecasting and evaluation have benefited from this and so have I, not least because it has stimulated Murphy to spend more time thinking about these issues than anyone.

Blattenberger notes that econometric forecasting is a promising area for increased use of scoring rules. I agree and have noticed an increasing use of probabilities, both subjective probabilities from economic forecasters and model-based probabilities from econometric models, in forecasts of a variety of economic variables. Similarly, the field of expert systems and artificial intelligence is moving toward greater use of probabilities, as Ríos mentions, and scoring rules could play a role in improving models of this ilk. More generally, the expansion of practical application of Bayesian methods has perhaps the greatest potential in terms of size and scope for evaluating probabilities, but very little has been done in the way of actual evaluation of probabilities from Bayesian models via scoring rules.

Terminology. Oliver notes the potential confusion between two uses of the term "scores." There are yet other uses of this common term, all of which probably have valid historical antecedents, and I can only hope

that when reports of the use of scoring rules are written, readers will take a careful enough look to understand what is being done!

In conclusion, let me thank the discussants once again, both for their kind words about some aspects of my paper and for their valuable suggestions for future directions. I hope that more work on probability forecasting and on the evaluation of probabilities as well as more applications of these topics in practice might be stimulated by this exchange, and I thank José Bernardo for making it possible.

ADDITIONAL REFERENCES IN THE DISCUSSION

Berger, J. (1994). An overview of robust Bayesian analysis. *Test* **3**, 5–124 (with discussion).

Berger, J. O. and Wolpert, R. L. (1984). *The Likelihood Principle*. Lecture notes-monograph series. IMS: Hayward.

Bernardo, J. M. (1979). Expected information as expected utility. *Ann. Statist.* **7**, 686–690.

Bernardo, J. M. (1987). Approximations in statistics from a decision-theoretical viewpoint. *Probability and Bayesian Statistics* (R. Viertl, ed.). New York: Plenum, 53–60.

Blattenberger, G. (1996). Money demand revisited: an operational subjective approach, *J. Appl. Econometrics* **11**, 153–168

Blattenberger, G. and Lad, F. (1988). An application of operational-subjective statistical methods to rational expectations, *J. Bus. Econ. Statistics* **6**, 453–477 (with discussion).

Cervera, J. L. and Muñoz, J. (1996). Proper scoring rules for fractiles. *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press.

Chaloner, K., Church, T., Louis, T. and Matts, J. (1993). Graphical elicitation of a prior distribution for a clinical trial. *The Statistician* **41**, 342–353.

Cooke, R. (1991). *Experts in Uncertainty*. Oxford: University Press.

Dawid, A. P. (1986). Probability forecasting. *Encyclopedia of Statistical Sciences* **7** (S. Kotz, N. L. Johnson and C. B. Read, eds.). New York: Wiley, 210–218.

Dawid, A. P., DeGroot, M. H. and Mortera, J. (1995). Coherent combination of experts' opinions. *Test* **4**, 263-313 (with discussion).

de Finetti, B. (1963). La décision et les probabilitiés. *Rev. Roumaine Math. Pures Appl.* **7**, 405–413.

de Finetti, B. (1964). Probabilità subordinate e teoria delle decisioni. *Rendiconti Matematica* **23**, 128–131. Reprinted as 'Conditional probabilities and decision theory' in 1972, *Probability, Induction and Statistics* New York: Wiley, 13–18.

Eaton, M. L. (1992). A statistical diptych: admissible inferences, recurrence of symmetric Markov chains. *Ann. Statist.* **20**, 1147–1179.

Edwards, W. and von Winterfeldt, D. (1986). *Decision Analysis and Behavioral Research*. Cambridge: University Press.

Fudenberg, D. and Tirole, J. (1991). *Game Theory*. Cambridge: University Press.

Hadley, G. and Kemp, M. C. (1971). *Variational Methods in Economics*. Amsterdam: North-Holland.

Harsanyi, J. (1967). Games with incomplete information played by 'Bayesian' players. *Manag. Sci.* **14** 159–182; 320–334; 486–502.

Hirshleifer, J. and Riley, J. G. (1992). *The Analytics of Uncertainty and Information*. Cambridge: University Press.

Kadane, J. B. (1993). Several Bayesians: a review. *Test* **2**, 1–32.

Kadane, J. B., Dickey, J. M., Winkler, R. L., Smith, W. S. and Peters, S. C. (1980). Interactive elicitation of opinion for a normal linear model. *J. Amer. Statist. Assoc.* **75**, 845–854.

Katz, R. W., Murphy, A. H. and Winkler, R. L. (1982). Assessing the value of frost forecasts to orchardists: A dynamic decision-making approach. *J. Appl. Meteor.* **21**, 518–531.

Krzysztofowicz, R. (1992). Bayesian correlation score: A utilitarian measure of forecast skill. *Mon. Wea. Rev.* **120**, 208–219.

Lindley, D. V. (1956). On a measure of information provided by an experiment. *Ann. Math. Statist.* **27**, 986–1005.

Lindley, D. V. (1982). Scoring rules and the inevitability of probability. *Internat. Statist. Rev.* **50**, 1–26 (with discussion).

McCloskey, D. and Ziliak, S. (1996). The standard error of regressions, *J. Economic Literature* **34**(1), 97-114,

Murphy, A. H. (1970). The ranked probability score and the probability score: A comparison. *Mon. Wea. Rev.* **98**, 917–924.

Murphy, A. H. (1991). Forecast verification: Its complexity and dimensionality. *Mon. Wea. Rev.* **119**, 1590–1601.

Murphy, A. H. (1995). A coherent method of stratification within a general framework for forecast verification. *Mon. Wea. Rev.* **123**, 1582–1588.

Murphy, A. H. (1996). Forecast verification. *Economic Value of Weather and Climate Forecasts* (R. W. Katz and A. H. Murphy, eds.). Cambridge: University Press, (to appear).

Murphy, A. H. and Daan, H. (1984). Impacts of feedback and experience on the quality of subjective probability forecasts: Comparison of results from the first and second years of the Zierikzee experiment. *Mon. Wea. Rev.* **112**, 413–423.

Murphy, A. H. and Ehrendorfer, M. (1996). *Probability forecasting and probability forecasts*. Corvallis, Oregon: Prediction and Evaluation Systems (manuscript).

Murphy, A. H. and Wilks, D. S. (1996). Statistical models in forecast verification: A case study of precipitation probability forecasts. *13th Conference on Probability and Statistics in the Atmospheric Sciences*. American Meteorology Society, 218–223.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Mateo: Morgan Kaufmann.

Pratt, J. W. and Zeckhauser, R. J. (eds.) (1985). *Principals and Agents: The Structure of Business*. Boston: Harvard Business School Press.

Rubin, H. (1987). A weak system of axioms for 'rational' behavior and the non-separability of utility from prior. *Statistics and Decisions* **5**, 47–58.

Schervish, M. J. (1995). *Theory of Statistics*. New York: Springer.

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L. and Cowell, R. G. (1993). Bayesian analysis in expert systems. *Statist. Sci.* **8**, 219–246.

Staël von Holstein, C. -A. S. and Murphy, A. H. (1978). The family of quadratic scoring rules. *Mon. Wea. Rev.* **106**, 917–924.

West, M. (1988). Modelling expert opinion. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 493–508 (with discussion).

Winkler, R. L. (1986). Expert resolution. *Manag. Sci.* **32**, 298–303.

Winkler, R. L., Smith, W. S. and Kulkarni, R. B. (1978). Adaptive forecasting models based on predictive distributions. *Manag. Sci.* **24**, 977-986.

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