MANAGEMENT SCIENCE
Vol. 29, No. 4, April 1983
Printed in U.S.A.

EFFECTIVE SCORING RULES FOR PROBABILISTIC FORECASTS*

DANIEL FRIEDMAN†

This paper studies the use of a scoring rule for the elicitation of forecasts in the form of probability distributions and for the subsequent evaluation of such forecasts. Given a metric (distance function) on a space of probability distributions, a scoring rule is said to be *effective* if the forecaster's expected score is a strictly decreasing function of the distance between the elicited and "true" distributions. Two simple, well-known rules (the spherical and the quadratic) are shown to be effective with respect to suitable metrics. Examples and a practical application (in Foreign Exchange rate forecasting) are also provided.
(FORECASTING, DELPHI TECHNIQUE)

1. Introduction

For as long as "experts" have made forecasts—that is, at least since the days of the Oracle at Delphi in ancient Greece—the "decision makers" who use these forecasts have had to face two problems. First, there is the elicitation problem: a decision maker needs assurance that it is in the expert's best interest to produce a forecast (the "true" forecast) that accurately reflects the expert's best judgment.[1] Second, there is the evaluation problem: a decision maker needs some way after the fact of judging the quality of the forecast he has received. The device of a *scoring rule*, that is, a rule which assigns to the expert some score based on his forecast of an event and on the actual outcome, evidently deals directly with the evaluation problem. However, insofar as the expert understands how he is to be evaluated and attempts to maximize his expected score, a scoring rule also addresses the elicitation problem. (See Murphy and Winkler [7] for a more careful discussion of these themes.)

A theoretical literature on scoring rules, based on the Savage–DeFinetti theory of personal probability, emerged in the 1960s. Scoring rules for probabilistic forecasts of dichotomous random variables (e.g., rain or no rain) were extended to cover expectations and distributions of random variables with a finite number of possible outcomes. See Savage [9] for an excellent exposition of this case as well as a discussion of scoring rules in contexts other than forecasting. A further extension to the case we study in this paper, in which the set of possible outcomes is a continuum (e.g., amount of rainfall), is presented in Matheson and Winkler [6].

In this literature, *proper* scoring rules—rules for which the expert can maximize his expected score by reporting his (subjectively) "true" forecast—are emphasized. There is a recurring theme in the literature that scoring rules should also have some stronger *monotonicity* property, so that "it pays . . . to keep any unavoidable discrepancy [between reported and "true" forecasts] small" (Savage [9, p. 787]). Such discrepancies are particularly difficult to avoid when one elicits the distribution function (or

*Accepted by Ambar G. Rao; received May 7, 1981. This paper has been with the author 3 months for 1 revision.

†University of California at Los Angeles.

[1] In the absence of a well-chosen incentive structure, the experts may indulge in game playing which distorts their stated probability distributions. For instance, casual observation of economic forecasters suggests that experts who feel they have a reputation to protect will tend to produce a forecast near the consensus, and experts who feel they have a reputation to build will tend to overstate the probabilities of events they feel are understated in consensus.

0025-1909/83/2904/0447$01.25

probability density) of a continuous random variable, because the set of such distributions is infinite dimensional, and it is not generally possible to specify precisely an arbitrary member of such a set. In practice, one would ask the forecaster to specify some member of a low-dimensional subset[2] of "admissible" distributions (e.g., a member of the two-dimensional set of normal distributions).

In this context, the use of a proper scoring rule provides no guarantee that the elicited distribution will be appropriate, because the "true" distribution will generally not be admissible. One requires the stronger property that the expected score is higher when the elicited distribution is closer to the "true" distribution ("closeness" being defined in terms of some appropriate distance function, i.e., metric). Scoring rules with this monotonicity property will be referred to below as *effective*.

The next section provides more precise definitions of these concepts, some examples of metrics on sets of density functions, and examples of scoring rules. §3 demonstrates the effectiveness of two of these scoring rules. The following section discusses practical application of the theory developed here. The final section summarizes our main argument, discusses briefly the related concept of "sensitivity to distance" (Stael von Holstein [10]), and points out generalizations and directions for further work.

2. Scoring Rules and Metrics

Let X be a random variable with distribution F and range $\Omega \subset R$. For the most part we will think of Ω as the set of possible realizations of X, and we will assume that F has a density function f belonging to some given set D of "conceivable" density functions defined on Ω.

A *metric*, or distance function, on D assigns to any pair f, g of densities in D a real number $d(f, g)$, which should be thought of as the "shortest distance" between f and g. By definition (see Rudin [8] for instance), a metric must satisfy the following three conditions for all f, g, h in D:

(a) positivity: $d(f, g) \geqslant 0$, and $d(f, g) = 0$ if and only if $f = g$;

(b) symmetry: $d(f, g) = d(g, f)$;

(c) triangle inequality: $d(f, g) \leqslant d(f, h) + d(h, g)$.

(The first two conditions are probably intuitively obvious requirements; the third says in essence that one cannot find a shorter route from f to g by first going to some h.)

A *scoring rule* is a real valued functional $S(g, x)$ defined for all g in D. S is (*strictly*) *proper* if its f-expected value, $E_f S(g) = \int S(g, x) f(x)\,dx$, is (strictly) maximized on D at $g = f$. It is *effective* with respect to d if, for all densities, f, g, and h in D, $E_f S(g) > E_f S(h) \Leftrightarrow d(f, g) < d(f, h)$. In words, the expected score is a monotone decreasing function of the distance between the true and elicited distributions. Note that if S is effective, then S is also strictly proper.

Some examples are now in order.

(a) The L_1-metric d_1 is defined by $d_1(f, g) = \int_\Omega |f(x) - g(x)|\,dx$. Geometrically, d_1 measures the area between the graphs of f and g. It is a continuous analogue of the mean absolute deviation between discrete distributions, $n^{-1}\sum_{i=1}^{n} |p_i - q_i|$, where $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_n)$ are the vectors of probabilities for a set of n discrete possible outcomes.

(b) The L_2-metric d_2 is defined by $d_2(f, g) = \{\int_\Omega |f(x) - g(x)|^2\,dx\}^{1/2}$. It is an ana-

[2] In fact, the set of admissible responses may be discrete (dimension 0) if forecasters uniformly "round off" their data. See De Finetti [1] for discussion of this point in the context of eliciting a probability distribution defined over a finite set of alternatives.

logue of the standard deviation (i.e., root mean squared deviation)

$$\left(n^{-1}\sum_{i=1}^{n}(p_i-q_i)^2\right)^{1/2}$$

between discrete distributions p and q. It also may be regarded as a generalization of the vector distance $(\sum_{i=1}^{n}(p_i-q_i)^2)^{1/2}$ between p and q. As such, it is a natural candidate for a metric on D.

(c) The *renormalized L_2-metric* d^* is defined as follows. Let $\|f\|_2=\{\int_\Omega|f(x)|^2\,dx\}^{1/2}$ be the L_2-norm of f, and $\rho f=f/\|f\|_2$ be the renormalization of f (i.e., $\rho f(x)=f(x)/c$, where $c=\|f\|_2$). Then $d^*(f,g)=d_2(\rho f,\rho g)$. The idea is that one rescales the densities so that they have unit L_2-norm before taking the L_2 distance between them. Proposition 3 of the next section proves that d^* is indeed a metric.

The reader may wish to compare these metrics by reference to some specific density functions. For instance, let

$$f(x)=\begin{cases}10 & \text{if } 0\leqslant x<0.10,\\ 0 & \text{otherwise;}\end{cases}\qquad g(x)=\begin{cases}10 & \text{if } 0.06\leqslant x<0.16,\\ 0 & \text{otherwise;}\end{cases}$$

and

$$h(x)=\begin{cases}0.01 & \text{if } 100\leqslant x<200,\\ 0 & \text{otherwise.}\end{cases}$$

One may readily compute $d_1(f,g)=\int_0^{0.06}10+\int_{0.10}^{0.16}10=10(0.12)=1.2$, while $d_1(f,h)=d_1(g,h)=2$. Hence, according to d_1, f and g are closer to each other than either is to h. Now $d_2(f,g)=(\int_0^{0.06}10^2+\int_{0.10}^{0.16}10^2)^{1/2}=\sqrt{12}$. Note that $\|f\|_2=\|g\|_2=\sqrt{10}$ and $\|h\|_2=0.1$, so

$$d_2(f,h)=\sqrt{\|f\|_2^2+\|h\|_2^2}=\sqrt{10.01}$$

and $d_2(g,h)=d_2(f,h)$. Hence, according to d_2, f and g are both closer to h than they are to each other (!). Finally,

$$d^*(f,g)=\left\{\int_0^{0.06}\left(10/\sqrt{10}\right)^2+\int_{0.10}^{0.16}\left(10/\sqrt{10}\right)^2\right\}^{1/2}=\sqrt{1.2}\,,$$

while

$$d^*(g,h)=d^*(f,h)=\left(\|f\|_2^2/\|f\|_2^2+\|h\|_2^2/\|h\|_2^2\right)^{1/2}=\sqrt{2}\,.$$

Hence, according to d^*, f and g are closer to each other than either is to h.

This example was chosen so that the densities had a form of the sort discussed in §4 and so that the distances were easy to compute. However, the example is suggestive in that informal experiments, involving the presentation to subjects of graphs of various density functions and asking which seem closest, have persuaded the author that d_1 and d^* both correspond rather closely with most subjects' intuitive idea of distance between functions, while d_2 (despite its statistical and Euclidean pedigrees) does not in general.

We turn now to examples of scoring rules.

(a) The naive scoring rule $N(g,x)=g(x)$ simply uses the height of density function at the realized outcome as the score. Despite its intuitive appeal, this rule is known to be improper, and therefore is not effective with respect to any metric. It is sometimes referred to in the literature as the "linear" rule; see Stael von Holstein [10, p. 148] for instance.

(b) Scoring rules are commonly derived from a decision maker's loss function. For instance, if the loss function is $l(a, x) = (a - x)^2$, where a is the "action" taken (i.e., the value assigned to the random variable) and x is the ex post value, then it is easy to see that one minimizes g-expected loss by setting $a = \bar{g}$, the mean of the elicited distribution. Thus, an appropriate scoring rule in this case would be $S(g,x) = -(\bar{g} - x)^2$. Since any g whose mean coincides with the "true" mean $\bar{f}$ receives the maximum score, this last rule is proper but not strictly proper, and hence not effective. See Murphy and Winkler [7, p. 284ff] for further discussion.

(c) The logarithmic scoring rule $L(g,x) = \log g(x)$ arises from information theory and can be shown to be strictly proper. However, it heavily penalizes underestimation of low probabilities and in fact gives an expected score of $-\infty$ to any g which is zero on a subset of support(f) of positive measure. If support(f) $\equiv \{x \mid f(x) > 0\}$ is not known a priori by the forecaster, practical difficulties arise. In Friedman [3] it is argued that there is reason to believe that L is not effective with respect to any metric, and that if this conjecture is true, it is a blow against maximum likelihood methods in statistical estimation. On the other hand, see Stael von Holstein [10, pp. 150–151] for a summary of the attractive features of L.

(d) The quadratic scoring rule $Q(g,x) = 2g(x) - \|g\|_2^2$ is again well known to be strictly proper. In the next section we will show that it is effective with respect to d_2.

(e) The spherical scoring rule $S(g,x) = g(x)/\|g\|_2$, like the quadratic, may be regarded as a "correction" to the naive rule. Again, it is well known to be strictly proper. It is effective with respect to d^*, as we will establish below.

3. Proofs of Effectiveness

This section will employ some basic techniques from functional analysis, so a few definitions are in order.[3] Let Ω be some measurable subset of R, perhaps R itself, and let $f: \Omega \to R$. Then the p-norm of f is

$$\|f\|_p = \left\{ \int_\Omega |f|^p \, dx \right\}^{1/p}, \qquad 1 \leqslant p < \infty.$$

The space of *p-integrable functions* is $L_p = \{f \mid \|f\|_p < \infty\}$. The usual metric on L_p is $d_p(f, g) = \|f - g\|_p$. The *unit sphere* in L_p is $B_p = \{f \in L_p \mid \|f\|_p = 1\}$. In the case $p = 2$ we have the *inner product* $(f, g) = \int_\Omega f(x)g(x)\,dx$, which will always be finite for $f, g \in L_2$. Finally, let $D = \{$bounded continuous density functions on $\Omega\}$; note that $D \subset B_1 \cap L_2$.

Proposition 1.[4] *The quadratic scoring rule $Q(g,x)$ is effective on D with respect to the L_2-metric d_2.*

Proof. First note that for any $f, g \in L_2$, $(d_2(f, g))^2 = (f - g, f - g)$, and $E_f Q(g) = 2(f, g) - (g, g)$. Therefore, for any $f \in L_2$ and $g, h \in D$,

$$\begin{aligned} d_2(f, g) < d_2(f,h) &\Leftrightarrow (f - g, f - g) < (f - h, f - h) \\ &\Leftrightarrow (f, f) + (g, g) - 2(f, g) < (f, f) + (h,h) - 2(f,h) \\ &\Leftrightarrow 2(f, g) - (g, g) > 2(f,h) - (h,h) \\ &\Leftrightarrow E_f Q(g) > E_f Q(h). \quad \text{Q.E.D.} \end{aligned}$$

The next three propositions develop the spherical scoring rule as a "correction" to the naive scoring rule, and show that it is effective with respect to the renormalized L_2-metric d^*.

[3] Rudin [8] is a handy reference for this material, but virtually any Real Analysis text would do.

[4] De Finetti [1] stated a discrete version of this proposition.

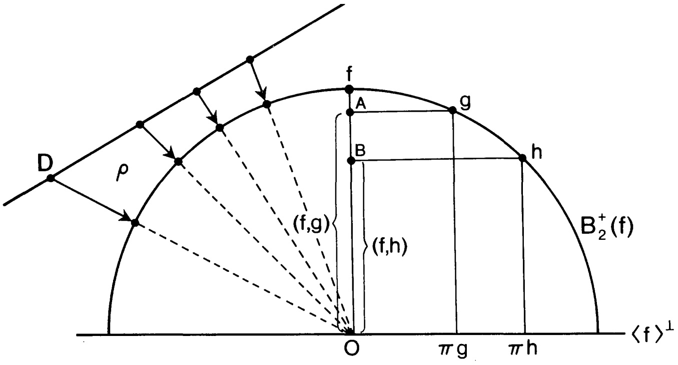

FIGURE 1. Diagrams for Propositions 2 and 3.

PROPOSITION 2. *For any $f \in B_2$, the naive scoring rule $N(g,x) = g(x)$ achieves a unique maximum in f-expected value on B_2 at $g = f$. Furthermore, $E_fN(g) > E_fN(h)$ $\Leftrightarrow d_2(f, g) < d_2(f,h)$ for all $g,h \in B_2^+(f) \equiv \{ g \in B_2 | (f, g) \geqslant 0\}$.*

PROOF. $E_fN(g) \equiv (f, g) \leqslant \|f\|_2 \cdot \|g\|_2 = 1$, by the well-known Schwarz inequality, with equality holding iff $f = g$, thus establishing the first part of the proposition.

For the second part, fix $f \in B_2$ and let $\pi : L_2 \rightarrow L_2$ be defined by $\pi g \equiv g - (f, g)f$ (orthogonal projection onto $\langle f\rangle^{\perp}$). The right-hand side of Figure 1 illustrates the argument we are about to make. We analyze the distance between f and g in terms of their distances from the point labeled "A," using the Pythagorean Theorem and likewise for the distance between f and h in terms of "B." To this end, pick $g,h \in B_2^+(f)$ and note that (by Pythagoras) $1 = \|g\|_2^2 = \|\pi g\|_2^2 + (f, g)^2\|f\|_2^2$, so $\|\pi g\|_2^2 = 1 - (f, g)^2$. Note also that $f - g = -\pi g + [1 - (f, g)]f$, and therefore $\|f - g\|_2^2 = \|\pi g\|_2^2 + [1 - (f, g)]^2\|f\|_2^2 = 2 - 2(f, g)$. Similarly, $\|f - h\|_2^2 = 2 - 2(f,h)$. The equivalences $E_fN(g) > E_fN(h) \Leftrightarrow (f, g) > (f,h) \Leftrightarrow \|f - g\|_2 < \|f - h\|_2 \Leftrightarrow d_2(f, g) < d_2(f,h)$ are now immediate. Q.E.D.

Proposition 1 shows that the naive scoring rule would be effective (with respect to d_2) if $D \subset B_2$, which is unfortunately not the case. However, D is contained in L_2 and $0 \notin D$, so we can project D onto B_2 by the spherical projection map ρ, where $\rho g \equiv g / \|g\|_2$. The left-hand side of Figure 1 illustrates ρ: the points in D at the tail of each arrow are mapped by ρ into the points in B_2 at the tip of the corresponding arrow. This map induces the renormalized L_2-metric d^* on D as the following proposition shows:

PROPOSITION 3. *The function $d^*(f, g) = d_2(\rho f, \rho g)$ is a metric on D.*

PROOF. $d^*(f, g) \geqslant 0$ is immediate from the definition. Suppose $d^*(f, g) = 0$. Then, by definition of L_2-norm, $\rho f = \rho g$ almost everywhere, i.e., $f(x) = c \cdot g(x)$ for almost every x, where the constant c is $\|f\|_2 / \|g\|_2$. Since f and $g \in D$, we can integrate both sides of the last equality and conclude $c = 1$ and therefore $f = g$. The symmetry of d is obvious, so there remains to establish only the triangle inequality $d^*(f, g) \leqslant d^*(f,h) + d^*(h, g)$. But this is an immediate consequence of the well-known Minkowsky inequality for the L_2-norm. Q.E.D.

If the mapping ρ is applied to the naive scoring rule, the spherical scoring rule results, i.e., $S(g,x) = N(\rho g,x)$. Since ρ always maps D into $B_2^+(f)$, we have the desired result:

PROPOSITION 4. *The spherical scoring rule $S(g,x) = g(x)/\|g\|_2$ is effective with respect to the metric d^*.*

Propositions 1 and 4 may easily be generalized in several ways. Note first that the set D of admissible density functions can be any subset of the bounded continuous densities defined on Ω. In turn, Ω can be any measurable subset of virtually any vector space, not just R. At the cost of some technical complication, D can represent probability distributions which do not necessarily have bounded and continuous density functions. The methods used above would seem valid as long as $L_2(\Omega)$ is dense in D. Therefore, the spherical and quadratic scoring rules are effective even if the random variable to be forecast is discrete, or bounded, or vector-valued, etc.

4. An Application[5]

The theory laid out in the previous two sections was originally developed in response to the following practical problem. The senior management of a multinational corporation requires forecasts of foreign exchange rates from its currency experts. The forecasts are input to a not fully specified but large and varied set of decisions, some of which require more detail than "point" (i.e., expected value) forecasts. The format requested for the forecasts is that of a "histogram" (i.e., piecewise constant density function) with five or fewer steps. The general form of the forecast is:

$$h(x) = \begin{cases} h_k & \text{if } \ x \in I_k, \quad k = 1, \dots, 5, \\ 0 & \text{otherwise,} \end{cases}$$

with $\sum_{i=1}^{5} p_i = 1.00$, where $p_i = w_i h_i$ and $w_i = \text{width } I_i$, the intervals I_i adjoining. The set A of admissible forecasts consists here of all densities of this form, i.e., each forecast is a message consisting of 10 numbers which specify the parameters p_k and I_k.

Having made such a request, management must deal with the elicitation and evaluation problems discussed in the introduction, and a scoring rule for density functions seems appropriate. Since a "loss function" cannot be specified in this case, management should consider strictly proper scoring rules. Moreover, since a forecaster's "true" subjective probability distribution, based on his best analysis of all information available to him, would not generally belong to the histogram family A, special consideration should be given to effective scoring rules.

Under the naive scoring rule, the forecaster receives the score $N(h,x) = h(x) = h_k = p_k / w_k$ if the actual exchange rate x falls into the forecast interval I_k. Thus he is rewarded both for a tight forecast (narrow interval) and an accurate one (high probability interval). To the extent that the forecaster attempts to maximize his expected score, however, this rule will provide him incentive to overstate the probabilities of the perceived modal (most likely) outcomes and thus to understate his perceived uncertainty, as the reader may verify by numerical examples. Such bias can be eliminated if management employs the spherical scoring rule

$$S(h,x) = h_k \left(\sum_{i=1}^{5} h_i p_i \right)^{-1/2} = h_k \left(\sum_{i=1}^{5} h_i^2 w_i \right)^{-1/2}.$$

As we have seen, this will motivate the expected-score maximizing forecaster to produce a histogram forecast "as close as possible" to his true distribution, where we take "closeness" in the sense of the renormalized L_2 metric d^*. If, for some reason, the metric d_2 were thought to better represent closeness of approximation for densities, management could use the quadratic scoring rule $Q(h,x) = 2h_k - \sum_{i=1}^{5} h_i p_i = 2h_k - \sum_{i=1}^{5} h_i^2 w_i$.

[5] From 7/78 to 5/81, Bank of America used an approach to FX rate forecasting quite similar to that sketched in this section.

In theory, management's elicitation and evaluation problems have now been solved. In practice, two further considerations seem important. First, a forecaster may not actually attempt to maximize his expected score, perhaps because he is risk averse. In this case, he may overstate his perceived uncertainty. A rough but practical solution to this problem is to use a rule such as the naive one, whose bias acts in a direction opposite to that of risk aversion by the forecaster.[6] Second, and perhaps more important, the forecaster must fully understand the scoring rule if he is to be properly motivated by it. Since some forecasters have difficulty with logs and square roots, the naive rule has much to recommend itself from the standpoint of simplicity.

A useful compromise, which has worked well in actual practice, has been to compute both the naive and spherical scores, and to emphasize the naive score in "track record" summaries of forecaster performance.[7] The spherical score then served as a vital check to ensure that the relative naive scores did not provide false signals to the forecasters and management. It also can be used to compute (confidential) weights assigned to forecasters in computing composite forecasts.[8]

5. Discussion

The main argument may be summarized as follows. There are circumstances, illustrated in the previous section, in which one must elicit information from an expert in the form of a simplified probability distribution. To motivate the expert to provide a simplified distribution as close as possible to his "true" subjective distribution, one should use an effective scoring rule, and not a merely proper scoring rule. Effectiveness is defined relative to some idea of closeness, or metric; but there is no universally accepted metric for probability distributions. However, the metric d^* defined in §2 does seem to correspond to intuitive ideas of distance, and we have seen that the spherical scoring rule is effective with respect to d^*. An alternative well-known scoring rule, the quadratic, was shown to be effective with respect to the L_2-metric. Once an appropriate rule for elicitation is selected, the expert's scores can be compiled into "track records," providing a basis for evaluating his performance relative to that of other experts.

The concept of effectiveness introduced in this paper does have antecedents in the literature. As suggested in the introduction, it is a generalization to the probability distribution case of Savage's monotonicity property of scoring rules for the expectation case. In some ways, effectiveness is also related to the concept of "sensitivity to distance" introduced around 1970. Roughly speaking, if g places more of its probability mass near the actual outcome than does h, then g would receive the higher score if the rule is "sensitive to distance." See Stael von Holstein [10] for a precise definition and Epstein [2] for an example of such a scoring rule. It is clear that this notion and our notion of "effectiveness" both spring from the same intuitive motivation, but have quite different formalizations. The "sensitivity to distance" concept employs a partial ordering and is outcome dependent. Effectiveness is based on the cardinal concept of a metric, and hence employs a complete ordering. Moreover, it is *not* outcome depen-

[6] Winkler [11] provides a partial theoretical resolution to the risk aversion problem: if utility $U(S)$ is a known function of score, one can compensate for risk aversion by replacing a scoring rule S by $\hat{S} = U^{-1} \cdot S$; clearly if S is effective for an expected score maximizer, then $\hat{S}$ is effective for an expected-utility-of-score maximizer. Holt [4] has an interesting discussion of how U itself might be elicited. See Murphy and Winkler [7] for a broader examination of this issue.

[7] The "track record" should also be used to check the consistency of the forecasts. For instance, if fewer than 50% of the actual outcomes lie in the central 50% confidence interval of their histogram forecasts, then we have evidence of a bias towards underestimating uncertainty. This sort of evidence can provide valuable feedback to the expert and help him improve his performance.

[8] See Kabus [5] for a development of this idea.

dent, but rather involves a comparison of the forecast to the (subjective) "true" distribution. For this reason, rules which are "sensitive to distance" in the Stael von Holstein sense are not effective in general. See Matheson and Winkler [6, p. 1092] for a related discussion.

We believe that the work presented here provides an adequate basis for "real-life" application of effective scoring rules, and in particular points out some advantages of the spherical scoring rule which are not readily apparent. However, it at best merely opens the door to serious theoretical study of effectiveness. One might like to know which scoring rules are effective with respect to *some* metric, however exotic—the case of the logarithmic rule being particularly significant in this regard. Equally, one might like to know which metrics allow effective scoring rules—the case of the L_1-metric being especially significant here.[9] We have developed the concept of effectiveness here in terms of metrics on density functions, but for some purposes one might prefer to use the weaker notion of distance between cumulative distribution functions, and much work remains to be done in this area. Friedman [3] contains some tentative first steps toward answering these and related questions, but the field remains wide open.[10]

[9] An anonymous referee of this journal has pointed out that such scoring rules would be relatively easy to teach, due to the nice geometric interpretation of the L_1-metric.

[10] Without implicating them, the author would like to thank Ray Peters of Bank of America for suggesting the problem which led to this paper, Ted Groves for pointing out the relevant literature, and Joe Ostroy, Ed Leamer, Werner Ploberger, and Tom Rothenberg for helpful discussions as the work crystallized into its present form. Also, many thanks are due to a referee of this journal for extensive and thoughtful suggestions and comments.

References

1. De Finetti, B., "Methods for Discriminating Levels of Partial Knowledge Concerning a Test Item," *British J. Math. and Statist. Psych.*, Vol. 18 (May 1965), pp. 87–123.
2. Epstein, E. S., "A Scoring System for Probability Forecasts of Ranked Categories," *J. Appl. Meteorol.*, Vol. 18 (December 1969), pp. 985–987.
3. Friedman, D., "Effective Scoring Rules for Probability Distributions," UCLA, Department of Economics Discussion Paper No. 164, November 1979.
4. Holt, C. A., "Elicitation of Subjective Probability Distributions and von Neumann–Morgenstern Utility Functions," Department of Economics Discussion Paper No. 79-128, University of Minnesota, 1979.
5. Kabus, I., "You Can Bank on Uncertainty," *Harvard Bus. Rev.*, Vol. 54 (May 1976), pp. 95–105.
6. Matheson, J. E. and Winkler, R. L., "Scoring Rules for Continuous Probability Distributions," *Management Sci.*, Vol. 22 (1976), pp. 1087–1096.
7. Murphy, A. H. and Winkler, R. L., "Scoring Rules in Probability Assessment and Evaluation," *Acta Psych.*, Vol. 34 (1970), pp. 273–286.
8. Rudin, W., *Real and Complex Analysis*, McGraw-Hill, New York, 1966.
9. Savage, L. J., "Elicitation of Personal Probabilities and Expectations," *J. Amer. Statist. Assoc.*, Vol. 66 (1971), pp. 783-801.
10. Stael von Holstein, C.-A. S., "Measurement of Subjective Probability," *Acta Psych.*, Vol. 34 (1970), pp. 146–159.
11. Winkler, R. L., "Scoring Rules and the Evaluation of Probability Assessors," *J. Amer. Statist. Assoc.*, Vol. 64 (September 1969), pp. 1073–78.