MANAGEMENT SCIENCE
Vol. 31, No. 5, May 1985
Printed in U.S.A.

SHOULD SCORING RULES BE 'EFFECTIVE'?*

ROBERT F. NAU
A. B. Freeman School of Business, Tulane University,
New Orleans, Louisiana 70118

A scoring rule is a reward function for eliciting or evaluating forecasts expressed as discrete or continuous probability distributions. A rule is *strictly proper* if it encourages the forecaster to state his true subjective probabilities, and *effective* if it is associated with a metric on the set of probability distributions. Recently, the property of effectiveness (which is stronger than strict properness) has been proposed as a desideratum for scoring rules for continuous forecasts, for reasons of "monotonicity" in keeping the forecaster close to his true probabilities, since in practice the forecast must be chosen from a low-dimensional set of "admissible" distributions. It is shown in this paper that what effectiveness implies, beyond strict properness, is not a monotonicity property but a *transitivity* property, which is difficult to justify behaviorally. The logarithmic scoring rule is shown to violate the transitivity property, and hence is not effective. The L_1 and L_∞ metrics are shown to allow no effective scoring rules. Some potential difficulties in interpreting admissible forecasts are also discussed.

(PROBABILITY FORECASTING; EVALUATION OF FORECASTS; PROPER SCORING RULES; EFFECTIVENESS OF SCORING RULES)

1. Introduction

Consider a forecast for an uncertain quantity, given in the form of a probability distribution over the set of possible values. After the true value has become known, it may be desired to assign the forecast a numerical score, either as the basis for a monetary reward to the forecaster, or simply as a statistic by which to evaluate and compare different forecasts. A function which assigns a score to every possible combination of a probability forecast and a realized value for the uncertain quantity is known as a *scoring rule*. A generally accepted desideratum for a scoring rule is that it should lead the forecaster to state his true probabilities in order to maximize his subjective expected score—i.e., there should be no incentive to hedge. This condition is known as *properness*—or, if the maximization is always unique, *strict properness*. Savage (1971) shows that the class of proper scoring rules can be identified on one hand with the class of all convex functions on the set of possible forecasts, and on the other hand with the class of all statistical decision problems with respect to the uncertain quantity. Thus, the class of proper scoring rules is mathematically rich, and the elicitation of forecasts under proper scoring rules is in some sense a microcosm of the general problem of optimal decision-making under uncertainty. Savage's analysis is concerned with *expected-value* forecasts for uncertain vector quantities; a discrete *probability* forecast represents the special case of an expected-value forecast in which the set of possible values for the uncertain quantity is the set of unit vectors in $\mathbb{R}^n$. Matheson and Winkler (1976) demonstrate that a proper scoring rule for binary discrete probability forecasts can be used to generate a family of proper scoring rules for *continuous* probability forecasts—i.e. forecasts given in terms of continuous probability distributions—and a rigorous generalization of Savage's main result to the continuous case is given by Haim (1982).

A recent paper by Friedman (1983) proposes a stronger requirement for scoring rules for continuous forecasts, noting that " . . . the set of such distributions is infinite

*Accepted by Robert L. Winkler; received December 12, 1983. This paper has been with the author 2 months for 1 revision.

0025-1909/85/3105/0527$01.25

dimensional, and it is generally not possible to specify precisely an arbitrary member of such a set." Hence, "(i)n practice one would ask the forecaster to specify some member of a low-dimensional subset of 'admissible' distributions (e.g., a member of the two-dimensional set of normal distributions)." It is asserted that, "*(i)n 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).* [Emphasis added] Scoring rules with this monotonicity property will be referred to . . . as *effective*." (1983, p. 448) The concept of effectiveness illuminates some interesting properties of the quadratic and spherical scoring rules, which Friedman shows to be effective with respect to the L_2 metric and the "renormalized L_2 metric," respectively. The question is subsequently posed: "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." (1983, p. 454)

This paper will develop more fully the relationship between strictly proper scoring rules and metrics. It will be shown that what the requirement of effectiveness adds to that of strict properness is not a monotonicity property, but instead a transitivity property, which is difficult to justify behaviorally, and which is not satisfied by the logarithmic rule. Furthermore, it will be shown that neither the L_1 nor the L_∞ metric allows an effective scoring rule. These results suggest that the requirement of effectiveness is excessively restrictive without resolving the difficulties that may arise in articulating and interpreting continuous probability forecasts.

2. Preliminaries

This section introduces some basic notation and definitions for scoring rules and metrics which can be applied interchangeably to the discrete and continuous cases. We are concerned with probability forecasts for an uncertain quantity, x, whose value is to be drawn from a sample space, X. In the discrete case, X will be represented as a set of positive integers: $X = \{1, 2, \ldots, n\}$, where $n \leqslant \infty$. In the continuous case, X will be represented by a subset of the real numbers: $X \subseteq \mathbb{R}$. The symbols $\mathbf{f}$, $\mathbf{g}$, and $\mathbf{h}$ will be used to denote general probability vectors representing distributions on the sample space—either "external" forecasts or "internal" subjective probability distributions for the forecaster. (The distinction will be clear from the context.) We will refer to the L_k *norm* of a vector (a measure of its length):

$$|\mathbf{f}|_k \equiv \left(\sum_{x=1}^{n} |f_x|^k \right)^{1/k} \quad \text{or} \quad \left(\int_{-\infty}^{\infty} |f(x)|^k \, dx \right)^{1/k}.$$

The L_1 norm is simply the *sum* of the absolute values of the components of a vector in the discrete case, and the *integral* of the absolute value in the continuous case. The L_∞ norm (also known as the Tchebycheff norm) is the *maximum* of the absolute values of the components in the discrete case, or the supremum of the absolute value in the continuous case. The L_2 norm measures length in the ordinary Euclidean sense. A vector $\mathbf{f}$ is a probability distribution if it is nonnegative and satisfies $|\mathbf{f}|_1 = 1$. A *forecast* is defined here to be a probability distribution which is bounded, i.e. whose L_∞ norm is finite. In the continuous case, this means that a forecast may not concentrate a finite amount of probability mass on a single point.

The distance between two vectors in a set V can be measured in terms of a *metric*, which is a function $d\colon V \times V \to \mathbb{R}$ having the following properties:

(i) *positivity*: $0 \leqslant d(\mathbf{f}, \mathbf{g})$, with equality only if $\mathbf{g} = \mathbf{f}$;

(ii) *symmetry*: $d(\mathbf{f}, \mathbf{g}) = d(\mathbf{g}, \mathbf{f})$; and

(iii) *triangle inequality*: $d(\mathbf{f}, \mathbf{h}) \leqslant d(\mathbf{f}, \mathbf{g}) + d(\mathbf{g}, \mathbf{h})$.

(Strictly speaking, in the continuous case, a metric is defined on a set of equivalence classes of functions, where two functions $\mathbf{f}$ and $\mathbf{g}$ are considered equivalent if they agree almost everywhere.) A metric is usually defined in terms of the length of the difference between two vectors, under an appropriate norm. E.g., the L_k *metric* is defined as:

$$d_k(\mathbf{f}, \mathbf{g}) = |\mathbf{f} - \mathbf{g}|_k \qquad \text{for} \quad k = 1, 2, \ldots, \infty.$$

For any k, d_k is a metric on the set of vectors whose L_k norm is finite. We will also refer to the "renormalized L_2 metric," denoted d^*, which is obtained by applying the L_2 metric to the difference between two vectors after they have first been renormalized to unit L_2-length, i.e. projected onto the surface of the unit sphere:

$$d^*(\mathbf{f}, \mathbf{g}) = \left| \frac{\mathbf{f}}{|\mathbf{f}|_2} - \frac{\mathbf{g}}{|\mathbf{g}|_2} \right|_2 .$$

Friedman (1983) shows that d^* is a metric on the set of bounded, continuous probability distributions. The *indicator distribution* for the value x, which will be denoted $\mathbf{e}_x$, is defined as the distribution which places unit mass on x and zero mass (or density) on all other values. In the discrete case, $\mathbf{e}_x$ is simply the xth unit vector in $\mathbb{R}^n$, and in the continuous case $\mathbf{e}_x(z) = \delta(z - x)$, where δ is the Dirac delta function. (In the latter case, $\mathbf{e}_x$ may be manipulated as a "generalized function," as described by Lighthill 1958.)

It will be convenient to represent a *scoring rule* as a real-valued function S whose arguments are the forecast distribution and the indicator distribution for x, rather than the forecast and x itself. That is, $S(\mathbf{f}, \mathbf{e}_x)$ denotes the forecaster's score when his forecast is the distribution $\mathbf{f}$, and the uncertain quantity takes on the value x. The forecaster's *expected score* for the forecast $\mathbf{f}$ when his true distribution is $\mathbf{g}$ can then be simply represented as $S(\mathbf{f}, \mathbf{g})$, where:

$$S(\mathbf{f}, \mathbf{g}) \equiv \sum_{x=1}^{n} g_x S(\mathbf{f}, \mathbf{e}_x) \quad \text{or} \quad \int_{-\infty}^{\infty} f(x) S(\mathbf{f}, \mathbf{e}_x)\, dx. \tag{1}$$

Note that $S(\mathbf{f}, \mathbf{g})$ is a linear function of $\mathbf{g}$. Thus, we may consider a scoring rule to be any real-valued function of the forecast and the true distribution that is linear in the true distribution. It will also be convenient to use *inner product* notation, where the inner product of two vectors is defined as:

$$\langle \mathbf{f}, \mathbf{g} \rangle = \sum_{x=1}^{n} f_x g_x \quad \text{or} \quad \int_{-\infty}^{\infty} f(x) g(x)\, dx.$$

For example, by taking the inner product of a forecast with an indicator distribution, we select the probability mass or density assigned to a particular outcome: $\langle \mathbf{f}, \mathbf{e}_x \rangle = f_x$ or $f(x)$. Also, note that the L_2 norm can be expressed in terms of the inner product of a vector with itself: $|\mathbf{f}|_2 = \langle \mathbf{f}, \mathbf{f} \rangle^{1/2}$.

A scoring rule S is defined to be [*strictly*] *proper* if $S(\mathbf{f}, \mathbf{g}) \leqslant S(\mathbf{g}, \mathbf{g})$ for all $\mathbf{f}$ and $\mathbf{g}$ in C [with equality only if $\mathbf{g} = \mathbf{f}$]. Some common, strictly proper scoring rules which will be discussed below are the "quadratic" scoring rule:

$$S_Q(\mathbf{f}, \mathbf{e}_x) \equiv 2\langle \mathbf{f}, \mathbf{e}_x \rangle - \langle \mathbf{f}, \mathbf{f} \rangle,$$

the "spherical" scoring rule:

$$S_S(\mathbf{f}, \mathbf{e}_x) \equiv \frac{\langle \mathbf{f}, \mathbf{e}_x \rangle}{\langle \mathbf{f}, \mathbf{f} \rangle^{1/2}},$$

and the "logarithmic" scoring rule:

$$S_L(\mathbf{f}, \mathbf{e}_x) \equiv \log(\langle \mathbf{f}, \mathbf{e}_x \rangle).$$

These definitions apply to either the continuous or discrete case, under the appropriate definition of the inner product.

From a purely formal standpoint, if would suffice to consider only the continuous case, since it subsumes the finite and countable discrete cases. For example, every n-dimensional discrete distribution $\mathbf{f}$ can be associated with a piecewise uniform distribution $f(z)$, where $f(z) = f_i$ for $i - 1 < z \leqslant i$, and $f(z) = 0$ for $z \leqslant 0$ or $z > n$. The L_k norm of the vector $\mathbf{f}$ is then identical to the L_k norm of the corresponding function $f(z)$. Correspondingly, the discrete event x (where $x \in \{1, \ldots, n\}$) can be associated with the uniform distribution on $(x - 1, x]$, denoted by $u_x(z)$. Thus, a discrete event can always be associated with a partition of a continuous sample space, regardless of whether this is the way it actually arose. If S is a [strictly] proper continuous scoring rule whose domain includes the piecewise uniform distributions on $(0, n]$, then $S(f(z), u_x(z))$ is a [strictly] proper scoring rule for the forecast vector $\mathbf{f}$ with respect to the discrete event x. If $\mathbf{g}$ is the true discrete distribution, the appropriate *expected* score is given by $S(f(z), g(z))$, where $g(z)$ is the piecewise uniform distribution corresponding to $\mathbf{g}$. In particular, when these definitions for $f(z)$ and $u_x(z)$ are substituted in the expressions given above for the continuous versions of the quadratic, spherical, and logarithmic scoring rules (taking the continuous inner product), the corresponding discrete versions of these rules (taking the discrete inner product) are obtained.

There are several reasons, though, for presenting the discrete case side by side with the continuous case. First, it will be convenient to use *discrete* counterexamples to establish certain results for *both* the discrete and continuous cases. Second, in discussing the sample application given by Friedman, it will be shown that when a proper continuous scoring rule is applied to a piecewise uniform forecast $f(z)$ elicited under an "admissibility" restriction, the score is *not necessarily proper* for the corresponding discrete forecast $\mathbf{f}$ with respect to the implied partition of the sample space. In practice, therefore, it may be preferable to partition the sample space at the outset, and then explicitly use only discrete forecasts and scoring rules.

3. Necessary Conditions for Effectiveness and Co-effectiveness

A scoring rule S is defined to be *effective* if for some metric d:

$$S(\mathbf{h}, \mathbf{g}) < S(\mathbf{f}, \mathbf{g}) \Leftrightarrow d(\mathbf{f}, \mathbf{g}) < d(\mathbf{h}, \mathbf{g})$$

for $\mathbf{f}$, $\mathbf{g}$, and $\mathbf{h}$ in C. A metric will be said to be *co-effective* if there exists a scoring rule which is effective with respect to it. From the positivity property of a metric, it follows that a scoring rule is strictly proper if it is effective. Friedman (1983) shows that the quadratic scoring rule is effective with respect to the L_2 metric, and the spherical scoring rule is effective with respect to the renormalized L_2 metric.

Now, a significant property of a scoring rule (regardless of whether it is proper) is that it is by definition a linear function of one of its arguments. This imposes the following necessary condition on co-effective metrics:

PROPOSITION 1. *Under a co-effective metric, if* **a** *is equidistant from* **f** *and* **g**, *and* **b** *is also equidistant from* **f** *and* **g**, *then every convex combination of* **a** *and* **b** *is equidistant from* **f** *and* **g**.

PROOF. Suppose that $\mathbf{a}$ and $\mathbf{b}$ are each equidistant from $\mathbf{f}$ and $\mathbf{g}$ under the metric d—i.e., $d(\mathbf{f},\mathbf{a}) = d(\mathbf{g},\mathbf{a})$ and $d(\mathbf{f},\mathbf{b}) = d(\mathbf{g},\mathbf{b})$. If d is co-effective, then for some scoring rule S it follows that $S(\mathbf{f},\mathbf{a}) = S(\mathbf{g},\mathbf{a})$ and $S(\mathbf{f},\mathbf{b}) = S(\mathbf{g},\mathbf{b})$. If $\mathbf{c}$ is a convex combination of $\mathbf{a}$ and $\mathbf{b}$—i.e. $\mathbf{c} = r\mathbf{a} + (1-r)\mathbf{b}$ where $0 \leqslant r \leqslant 1$—it follows from the linearity property of scoring rules that:

$$S(\mathbf{f},\mathbf{c}) = rS(\mathbf{f},\mathbf{a}) + (1-r)S(\mathbf{f},\mathbf{b}) = rS(\mathbf{g},\mathbf{a}) + (1-r)S(\mathbf{g},\mathbf{b}) = S(\mathbf{g},\mathbf{c}).$$

Since S is effective with respect to d, this implies $d(\mathbf{f},\mathbf{c}) = d(\mathbf{g},\mathbf{c})$—i.e., $\mathbf{c}$ is equidistant from $\mathbf{f}$ and $\mathbf{g}$, as asserted.

This requirement can be shown to rule out both the L_1 and L_∞ metrics. Consider the 3-dimensional case: let $\mathbf{f} = (0.3, 0.3, 0.4)$, $\mathbf{g} = (0.45, 0.35, 0.2)$, $\mathbf{a} = (0.35, 0.35, 0.3)$, and $\mathbf{b} = (0.45, 0.2, 0.35)$. Under the L_1 metric, the distance between two vectors is the sum of the absolute differences in their respective components, whence: $d_1(\mathbf{f},\mathbf{a}) = 0.2$, $d_1(\mathbf{g},\mathbf{a}) = 0.2$, $d_1(\mathbf{f},\mathbf{b}) = 0.3$, and $d_1(\mathbf{g},\mathbf{b}) = 0.3$. I.e., $\mathbf{a}$ and $\mathbf{b}$ are each equidistant from $\mathbf{f}$ and $\mathbf{g}$ under d_1. Now let $\mathbf{c} = (\mathbf{a}+\mathbf{b})/2 = (0.4, 0.275, 0.325)$. Then $d_1(\mathbf{f},\mathbf{c}) = 0.2$, but $d_1(\mathbf{g},\mathbf{c}) = 0.25$. Thus, $\mathbf{c}$ is a convex combination of $\mathbf{a}$ and $\mathbf{b}$ but it is *not* equidistant from $\mathbf{f}$ and $\mathbf{g}$. The same result obtains under the L_∞ metric, which measures the *maximum* absolute difference in the components. In fact, for the 3-dimensional case, $d_\infty(\mathbf{f},\mathbf{g}) = d_1(\mathbf{f},\mathbf{g})/2$ if $\mathbf{f}$ and $\mathbf{g}$ are probability distributions. This example can also be used to disqualify these metrics in the higher-dimensional discrete case by appending further components which are all 0's, or in the continuous case by converting them into piecewise uniform functions, as noted above.

The symmetry property of metrics, on the other hand, implies what will be called the "transitivity property of effective scoring rules:"

PROPOSITION 2. *If S is an effective scoring rule, then*:
(i) $S(\mathbf{g},\mathbf{f}) \leqslant S(\mathbf{h},\mathbf{f})$ *and* $S(\mathbf{f},\mathbf{h}) \leqslant S(\mathbf{g},\mathbf{h}) \Rightarrow S(\mathbf{f},\mathbf{g}) \leqslant S(\mathbf{h},\mathbf{g})$,
(ii) $S(\mathbf{g},\mathbf{f}) < S(\mathbf{h},\mathbf{f})$ *and* $S(\mathbf{f},\mathbf{h}) \leqslant S(\mathbf{g},\mathbf{h}) \Rightarrow S(\mathbf{f},\mathbf{g}) < S(\mathbf{h},\mathbf{g})$,
(iii) $S(\mathbf{g},\mathbf{f}) = S(\mathbf{h},\mathbf{f})$ *and* $S(\mathbf{f},\mathbf{h}) = S(\mathbf{g},\mathbf{h}) \Rightarrow S(\mathbf{f},\mathbf{g}) = S(\mathbf{h},\mathbf{g})$.

PROOF. For part (i), assume $S(\mathbf{g},\mathbf{f}) \leqslant S(\mathbf{h},\mathbf{f})$ and $S(\mathbf{f},\mathbf{h}) \leqslant S(\mathbf{g},\mathbf{h})$. If S is effective with respect to a metric d, then the reverse inequalities hold for d: $d(\mathbf{g},\mathbf{h}) \leqslant d(\mathbf{f},\mathbf{h})$ and $d(\mathbf{h},\mathbf{f}) \leqslant d(\mathbf{g},\mathbf{f})$. But, by the symmetry property of metrics, we also have $d(\mathbf{f},\mathbf{h}) = d(\mathbf{h},\mathbf{f})$, whence $d(\mathbf{g},\mathbf{h}) \leqslant d(\mathbf{g},\mathbf{f})$. Invoking the symmetry property again to reverse the arguments on both sides, we obtain $d(\mathbf{h},\mathbf{g}) \leqslant d(\mathbf{f},\mathbf{g})$. Finally, since S is effective with respect to d, it follows that $S(\mathbf{f},\mathbf{g}) \leqslant S(\mathbf{h},\mathbf{g})$, as asserted. Parts (ii) and (iii) are proved similarly.

The transitivity property of effective scoring rules is illustrated by the following scenario: consider three experts (called F, G, and H) who have subjective probability distributions $\mathbf{f}$, $\mathbf{g}$, and $\mathbf{h}$, respectively, for some uncertain quantity. Suppose that their forecasts are simultaneously elicited under the same strictly proper scoring rule. Each expert will then reveal his own distribution; and, after all three distributions have been announced, each expert is able to compare the forecasts of the other two against his own *and against each other*, based on their expected scores according to his own distribution. If the scoring rule is not merely strictly proper but *effective*, then the transitivity property implies that if F feels H's forecast is better than G's, and H feels G's forecast is better than F's, it must follow that G will feel H's forecast is better than F's. That is, the three "between-others" comparisons must not form a cycle. While this would be a reasonable requirement for a triad of comparisons made by the same individual (say, a fourth expert), it is hard to justify where three different subjective viewpoints are represented.

For the logarithmic scoring rule, it is easy to generate examples of violations of the transitivity property. In the three-dimensional case, the logarithmic scoring rule is simply:

$$S_L(\mathbf{f},\mathbf{g}) = \sum_{i=1}^{3} g_i \log(f_i).$$

Letting $\mathbf{f} = (0.2, 0.4, 0.4)$, $\mathbf{g} = (0.26, 0.61, 0.13)$, and $\mathbf{h} = (0.5, 0.3, 0.2)$, we find:

$$S_L(\mathbf{h},\mathbf{f}) - S_L(\mathbf{g},\mathbf{f}) = 0.019, \quad S_L(\mathbf{g},\mathbf{h}) - S_L(\mathbf{f},\mathbf{h}) = 0.033 \quad \text{and}$$

$$S_L(\mathbf{f},\mathbf{g}) - S_L(\mathbf{h},\mathbf{g}) = 0.027,$$

which forms a cycle in violation of part (ii). This example, by extension, also disqualifies the logarithmic rule for higher-dimensional discrete forecasts or continuous forecasts. (If it is desired to avoid zero-probability pathologies—i.e. the possibility of an infinitely negative score—the 3 components of each vector could all be scaled down by, say, multiplying them by 0.99. The remaining probability mass of 0.01 could then be distributed, e.g. exponentially, over the remainder of the sample space in an identical fashion for all 3 forecasts. This would produce no significant change in the score differences given above.)

4. The Monotonicity Property of Strictly Proper Scoring Rules

Having established one property which effectiveness adds to that of strict properness —namely the transitivity property—we now turn to the question of whether effectiveness also implies a stronger "monotonicity" in some sense. First, note that a certain measure of *distance* between distributions is already implicit in the definition of a strictly proper scoring rule, and is defined by the "loss function" L, where: $L(\mathbf{f},\mathbf{g}) \equiv S(\mathbf{g},\mathbf{g}) - S(\mathbf{f},\mathbf{g})$. This function measures the expected loss incurred by the forecaster for announcing $\mathbf{f}$ rather than $\mathbf{g}$ as his forecast, when $\mathbf{g}$ is his true subjective probability distribution. We shall refer to a loss function as "strictly proper" or "effective" according to whether its corresponding scoring rule has those properties. If L is strictly proper, then, by definition, it satisfies the positivity property of a metric—i.e., $L(\mathbf{f},\mathbf{g}) \geqslant 0$, with equality only if $\mathbf{g} = \mathbf{f}$. (If it is proper but not strictly proper, then L is merely nonnegative.) The properties of a metric which are *not* generally satisfied by a strictly proper loss function are the symmetry property and the triangle inequality. In fact, Savage (1971, p. 788) shows that the only binary-event scoring rule for which the loss function is symmetric is the quadratic rule.

In sketching an historical perspective for the concept of effectiveness in the scoring-rule literature, Friedman (1983, p. 447) states: "In this literature, *proper* scoring rules . . . are emphasized. There is a recurring theme 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,'" where the inner quotation is from Savage (1971, p. 787). Later (p. 453), it is reiterated: "The concept of effectiveness does have its antecedents in the literature. As suggested in the introduction, it is a generalization to the probability case of Savage's monotonicity property of scoring rules for the expectation case." It will be argued here that Savage's monotonicity property applies to all strictly proper scoring rules for discrete or continuous probability forecasts, regardless of whether they are effective. The fuller text of Savage's remark is as follows: "The function L, and equivalently S, has an easily derived and useful monotonicity, according to which it not only pays to choose $\mathbf{f}$ equal to $\mathbf{g}$, but to keep any unavoidable discrepancy small. Namely, if $\mathbf{h}$ is between $\mathbf{g}$ and $\mathbf{f}$, then $L(\mathbf{h},\mathbf{g}) \leqslant L(\mathbf{f},\mathbf{g})$, with strict inequality if . . . [*strict* properness holds] and $\mathbf{f} \neq \mathbf{g}$." (Here our symbols $\mathbf{f}$, $\mathbf{g}$, $\mathbf{h}$, and S have been substituted for Savage's equivalents: $\mathbf{z}$, $\mathbf{x}$, $\mathbf{r}$, and I, respectively.) This assertion (and a constructive proof) appear in Savage's discussion of scoring rules for univariate expected-value forecasts, but the argument applies equally well to scoring rules for discrete or continuous probability forecasts, with the relation "**h** is between **g** and **f**" being replaced by its higher-dimensional generalization—i.e. the notion of a strictly convex combination:

PROPOSITION 3. *If S is proper, and if* **h** *is a strictly convex combination of* **f** *and* **g**, *then* $S(\mathbf{f},\mathbf{g}) \leqslant S(\mathbf{h},\mathbf{g})$, *and equivalently* $L(\mathbf{h},\mathbf{g}) \leqslant L(\mathbf{f},\mathbf{g})$. *If S is strictly proper, these relations hold with strict inequality.*

PROOF. Let **h** be a strictly convex combination of **f** and **g**—i.e., $\mathbf{h} = r\mathbf{f} + (1 - r)\mathbf{g}$, where $0 < r < 1$. If S is proper, then $S(\mathbf{f},\mathbf{h}) \leqslant S(\mathbf{h},\mathbf{h})$. Since S is linear in its second argument, this can be expanded as:

$$rS(\mathbf{f},\mathbf{f}) + (1 - r)S(\mathbf{f},\mathbf{g}) \leqslant rS(\mathbf{h},\mathbf{f}) + (1 - r)S(\mathbf{h},\mathbf{g})$$

which rearranges to:

$$(r/(1 - r))(S(\mathbf{f},\mathbf{f}) - S(\mathbf{h},\mathbf{f})) \leqslant S(\mathbf{h},\mathbf{g}) - S(\mathbf{f},\mathbf{g}).$$

If S is proper, the LHS of this inequality is nonnegative (or in the strictly proper case, positive), whence so is the RHS.

This proposition implies that $L(\mathbf{h},\mathbf{g})$ increases monotonically as **h** traverses a linear path (i.e. a continuously parameterized set of convex combinations) from **g** to **f**. Such "straight-line monotonicity" is a property which is not necessarily possessed by a metric. For example, consider a two-dimensional map of rugged 3-dimensional terrain. The minimum possible hiking distance between points on the map defines a metric on it, but traversing the path corresponding to a straight line on the map need not bring the hiker monotonically closer to his destination under this metric if he thereby goes over a tall and otherwise avoidable obstacle.

A co-effective metric for a continuous scoring rule, it will now be shown, must not only possess staight-line monotonicity, but must in fact be related to the loss function by a strictly increasing transformation. The exact form of this transformation, though, may depend on the point of reference—i.e. the forecaster's true distribution. In particular:

PROPOSITION 4. *A continuous scoring rule S is effective with respect to a metric d if-and-only-if for every* **g** *there is a strictly increasing function $t_{\mathbf{g}}$ which satisfies $t_{\mathbf{g}}(0) = 0$, such that*:

$$d(\mathbf{f},\mathbf{g}) = t_{\mathbf{g}}(L(\mathbf{f},\mathbf{g})) = t_{\mathbf{g}}(S(\mathbf{g},\mathbf{g}) - S(\mathbf{f},\mathbf{g})), \tag{2}$$

for all **f**.

PROOF. For the "if" part, note that the existence of such a function $t_{\mathbf{g}}$ would imply the following:

$$d(\mathbf{f},\mathbf{g}) < d(\mathbf{h},\mathbf{g}) \Leftrightarrow t_{\mathbf{g}}(L(\mathbf{f},\mathbf{g})) < t_{\mathbf{g}}(L(\mathbf{h},\mathbf{g}))$$

$$\Leftrightarrow L(\mathbf{f},\mathbf{g}) < L(\mathbf{h},\mathbf{g}) \Leftrightarrow S(\mathbf{h},\mathbf{g}) < S(\mathbf{f},\mathbf{g}),$$

establishing the effectiveness of S with respect to d. For the "only-if" part, we assume that S is continuous and effective with respect to a metric d, and must demonstrate the existence of an appropriate function $t_{\mathbf{g}}$. Since S is continuous, so is its loss function L. Since S is effective, it is strictly proper, and L therefore has the positivity property. Let $Z_{\mathbf{g}}$ denote the supremum of $L(\mathbf{h},\mathbf{g})$ as **h** ranges over the set of all forecasts, and note that $Z_{\mathbf{g}}$ must be positive. Since L is continuous, for any $z \in (0, Z_{\mathbf{g}})$ there must exist **h** such that $L(\mathbf{h},\mathbf{g}) = z$. Now, if $L(\mathbf{f},\mathbf{g}) = L(\mathbf{h},\mathbf{g})$ for two different forecasts **f** and **h**, then by effectiveness it follows that $d(\mathbf{f}, \mathbf{g}) = d(\mathbf{h}, \mathbf{g})$. Therefore, let $t_{\mathbf{g}}(z)$ be defined as the unique value of $d(\mathbf{h}, \mathbf{g})$ for all $\mathbf{h}$ such that $L(\mathbf{h}, \mathbf{g}) = z$. Since d and L share the positivity property, it follows that $t_{\mathbf{g}}(0) = d(\mathbf{g}, \mathbf{g}) = 0$. Next, let x and y satisfy $0 \leqslant x < y < Z_{\mathbf{g}}$. By the continuity of L, there exist vectors $\mathbf{h}_x$ and $\mathbf{h}_y$ such that $L(\mathbf{h}_x, \mathbf{g}) = x$ and $L(\mathbf{h}_y, \mathbf{g}) = y$. Since $L(\mathbf{h}_x, \mathbf{g}) < L(\mathbf{h}_y, \mathbf{g})$, it follows from the defintion of effectiveness that $d(\mathbf{h}_x, \mathbf{g}) < d(\mathbf{h}_y)$. From the definition of $t_{\mathbf{g}}$, we then have $d(\mathbf{h}_x, \mathbf{g}) = t_{\mathbf{g}}(x)$ and $d(\mathbf{h}_y, \mathbf{g}) = t_{\mathbf{g}}(y)$, whence $t_{\mathbf{g}}(x) < t_{\mathbf{g}}(y)$, which establishes that $t_{\mathbf{g}}$ is strictly increasing.

To demonstrate that a continuous scoring rule is effective with respect to a particular metric, it therefore suffices to demonstrate that a strictly increasing function $t_{\mathbf{g}}$ satisfying (2) exists for all $\mathbf{f}$ and $\mathbf{g}$. For the quadratic scoring rule, the loss function is $L(\mathbf{f}, \mathbf{g}) = |\mathbf{f} - \mathbf{g}|_2^2$, whence $t_{\mathbf{g}}(z) = z^{1/2}$. (Here, since $t_{\mathbf{g}}$ is independent of $\mathbf{g}$, the loss function shares the symmetry property of its associated metric, as noted by Savage.) For the spherical rule, $t_{\mathbf{g}}(z) = (2z/|g|_2)^{1/2}$. In fact, this approach may be used to establish the effectiveness of certain *families* of scoring rules based on the quadratic and spherical rules. For example, a generalized quadratic scoring rule, denoted S_Q^A, can be defined by: $S_Q^A(\mathbf{f}, \mathbf{e}_x) \equiv S_Q(A\mathbf{f}, A\mathbf{e}_x)$ where S_Q is the ordinary quadratic rule defined previously, and A is a linear operator. This rule is strictly proper if A is invertible (i.e. of "full rank") on the set of probability distributions. In the discrete case, A can be any nonsingular matrix (Stael von Holstein and Murphy 1978). In the continuous case, A can be a bounded nonzero weight function, or the differential operator, or the integral with respect to a probability measure. (The latter case is discussed by Matheson and Winkler 1976.) It is straightforward to show that the generalized quadratic scoring rule is effective with respect to the corresponding generalized L_2 metric: $d_2^A(\mathbf{f}, \mathbf{g}) \equiv d_2(A\mathbf{f}, A\mathbf{g})$, with $t_{\mathbf{g}}(z) = z^{1/2}$ as before. A similar generalization is possible for the spherical rule. Generalized quadratic scoring rules have been found to be useful in some applications—e.g. the so-called *ranked probability score*, in which A is a lower triangular matrix of 1's, is appropriate when discrete outcomes are meaningfully *ordered*, as in some weather forecasting situations. In general, the choice of a scoring rule should be tailored to the particular forecasting problem at hand, or undesirable, side effects may result. Against this possibility, effectiveness provides no guarantee, as will be illustrated below.

5. Considerations in Choosing an Admissible Set

In summary, it appears more correct to say that a co-effective metric is more monotonic than other metrics, rather than that an effective scoring rule is more monotonic than other (strictly proper) scoring rules. What an effective scoring rule possesses, beyond strict properness, is the transitivity property, whose merits are dubious. We should therefore reconsider the original motivating issue: admissibility. If a continuous forecast is to be drawn from a low-dimensional subset of admissible distributions, the forecaster's objective should be the same regardless of whether the scoring rule is effective or merely strictly proper: *a forecast should be selected which maximizes the expected score on the admissible set*. But the choice of a *particular* admissible set may determine the appropriateness of a given scoring rule. Several criteria to consider in this regard are: (i) how difficult is it for the forecaster to identify the admissible forecast which maximizes his expected score (i.e. what constrained maximization problem must he solve); and (ii) what inferences can be drawn by an outside observer about the forecaster's "true" probabilities, given his admissible forecast (i.e., what basis does it provide for betting or decision making by others)?

For example, in the application described by Friedman, the admissible set is the set of piecewise uniform distributions with five or fewer steps. Here, for any given five-fold partition of the continuum, the forecaster must choose the optimal assignment of probability mass to the five designated intervals—*and he must also choose the optimal partition.* This is a formidable mathematical programming problem to solve subjectively. Commonly used probability elicitation schemes often require much less: either a fixed partition is specified, so that a discrete forecast may be given, or else fractiles are specified, and the forecaster must only locate the corresponding intervals.

Besides the difficulty in articulating a piecewise uniform forecast, there is the problem of its interpretation. We might wish to be able to interpret the probability mass assigned to each interval as representing the forecaster's true probability for the outcome falling in that interval. As a simple example, suppose that the sample space is the unit interval, and that an admissible forecast may have only two steps—i.e. one break point. Suppose the forecast has its break point at $x = x^*$, with a probability mass of p (i.e. a constant density $f(x) = p/x^*$) assigned to the interval $(0, x^*]$, and a complementary probability mass of $1 - p$ (i.e. a constant density $f(x) = (1 - p)/(1 - x^*)$) on $(x^*, 1]$. It would seem reasonable to infer that p represents the forecaster's probability that $x \leqslant x^*$. This need not be the case, however, if the forecaster has attempted to maximize his expected score, even though the scoring rule may be effective. Consider the generalized quadratic scoring rule for which the operator A is a weight function—i.e. $Af(x) = a(x)f(x)$— and let $a(x)$ be defined as follows: $a(x) = 1$ for $1/4 < x \leqslant 3/4$, $a(x) = r^{1/2}$ otherwise, where $0 < r \leqslant 1$. Note that this is simply the ordinary (unweighted) quadratic scoring rule if $r = 1$. Now suppose that the forecaster's true distribution is the linear function $g(x) = 2x$. By simple calculus it can be shown that the optimal break point is always $x = 1/2$, but *the optimal density for each interval depends on* r—in particular, it is $f(x) = (3 + r)/(4 + 4r)$ for $x \leqslant 1/2$. Obviously, if $r \neq 1$, the optimal probability mass assigned to the interval $(0, 1/2]$ will not equal the forecaster's true probability mass for this interval, which is $1/4$.

In general, for a piecewise uniform forecast to reflect the forecaster's true assignment of probability mass among the intervals, it is necessary for the scoring rule to satisfy $S(\mathbf{f}, \mathbf{g}) = S(\mathbf{f}, \mathbf{g}^*)$ for every distribution $\mathbf{g}$ and every piecewise uniform forecast $\mathbf{f}$, where $\mathbf{g}^*$ denotes the piecewise uniform distribution having the same break points as $\mathbf{f}$ but assigning the same probability masses to the intervals as $\mathbf{g}$. This is true for both the unweighted spherical and quadratic rules since for them the expected score depends on $\mathbf{g}$ only through the inner product $\langle \mathbf{f}, \mathbf{g} \rangle$, which is unaffected when $\mathbf{g}$ is replaced by $\mathbf{g}^*$. (For these rules, expressions for the expected score may be obtained by substituting $\mathbf{g}$ for $\mathbf{e}_x$ in their definitions in §2.)

It seems likely that similar difficulties would arise if the admissible set were the family of normal distributions, and it is not clear under what scoring rules the optimal mean and variance would coincide with the forecaster's true mean and variance, if indeed there exist any scoring rules with this property. Thus, depending on the interpretation which is desired of the forecast, the choice of an admissible set can place particular demands on the scoring rule which are not addressed by mere considerations of effectiveness or strict properness.

References

FRIEDMAN, DANIEL, "Scoring Rules for Probabilistic Forecasts," *Management Sci.*, 29, 4 (April 1983), 447–454.

HAIM, E., "Characterization and Construction of Strictly Proper Scoring Rules," Ph.D. Thesis, University of California, Berkeley, 1982.

LIGHTHILL, M. J., *Fourier Analysis and Generalised Functions*, Cambridge University Press, London, 1958.

MATHESON, J. E. AND R. L. WINKLER, "Scoring Rules for Continuous Probability Distributions," *Management Sci.*, 22 (1976), 1087–1096.

SAVAGE, L. J. "Elicitation of Personal Probabilities and Expectations," *J. Amer. Statist. Assoc.*, 66 (1971), 783–801.

STAEL VON HOLSTEIN, CARL-AXEL S. AND ALLAN H. MURPHY, "The Family of Quadratic Scoring Rules," *Monthly Weather Rev.*, Vol. 106 (1978), 917–924.