A General Method for Comparing Probability Assessors

Mark J. Schervish

The Annals of Statistics, Vol. 17, No. 4 (Dec., 1989), 1856-1879.

Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28198912%2917%3A4%3C1856%3AAGMFCP%3E2.0.CO%3B2-A

The Annals of Statistics is currently published by Institute of Mathematical Statistics.

The Annals of Statistics
1989, Vol. 17, No. 4, 1856–1879

A GENERAL METHOD FOR COMPARING PROBABILITY ASSESSORS

BY MARK J. SCHERVISH

Carnegie Mellon University

A probability assessor or forecaster is a person who assigns subjective probabilities to events which will eventually occur or not occur. There are two purposes for which one might wish to compare two forecasters. The first is to see who has given better forecasts in the *past*. The second is to decide who will give better forecasts in the *future*. A method of comparison suitable for the first purpose may not be suitable for the second and vice versa. A criterion called calibration has been suggested for comparing the forecasts of different forecasters. Calibration, in a frequency sense, is a function of long run (future) properties of forecasts and hence is not suitable for making comparisons in the present. A method for comparing forecasters based on past performance is the use of scoring rules. In this paper a general method for comparing forecasters after a finite number of trials is introduced. The general method is proven to include calculating all proper scoring rules as special cases. It also includes comparison of forecasters in all simple two-decision problems as special cases. The relationship between the general method and calibration is also explored. The general method is also translated into a method for deciding who will give better forecasts in the future. An example is given using weather forecasts.

1. Introduction. In this paper we will consider a number of probability assessors (forecasters) who will assign probabilities to certain well-defined events. Intuitively, forecasts are good inasmuch as they are "close" to the indicators of the events being forecast. For example, a forecast of 0.1 for an event which does not occur will appear better, after the fact, than a forecast of 0.2 for that same event. In Section 2, we consider some existing methods used for comparing forecasters. In Section 3, we introduce a general method which includes the existing methods as special cases. The general method can be used either for comparing several forecasters or for evaluating a single forecaster. In Section 4, we characterize one of the existing methods, proper scoring rules, in terms of how it relates to the general method. A different characterization, not related to the general method, is given by Hendrickson and Buehler (1971). Some examples of the characterization are given in Section 5. In Sections 6 and 7, we show how other existing methods of comparing forecasters (calibration and dominance) relate to the general method. We examine the relationship between proper scoring rules and calibration in Section 8. An example, involving weather forecasting, of the use of the general method is given in Section 9.

A related question, not addressed in this paper, is how to improve a forecaster who is judged to be performing poorly according to the general method. Methods

Received January 1985; revised January 1989.

AMS 1980 *subject classifications*. Primary 62B15; secondary 62C10.

Key words and phrases. Calibration, dominance, forecasters, loss functions, refinement, scoring rules, sufficiency.

for improving forecasters have been considered by Lindley, Tversky and Brown (1979) and by Lindley (1982). These methods tend to rely on the use of beliefs about forecasters. In Section 10, we see how the general method for comparing forecasters can be translated into a general method for evaluating forecasters based on beliefs. It should be noted that none of the methods for comparing forecasters based on past performance is designed to determine which forecaster is likely to perform better in the future. However, it is hoped that the general method described in this paper can provide useful information for helping to decide who might perform better in the future.

Much literature exists on the subject of comparing forecasts in meteorology. An excellent survey of this area is given by Murphy and Winkler (1984). Some work has been done comparing probabilistic forecasts to categorical forecasts in decision problems by Thompson (1952, 1962), Thompson and Brier (1955), Murphy (1977) and Winkler and Murphy (1979). In this paper, we will consider only probabilistic forecasts and we will compare different forecasters to each other. Other authors have considered the question of the goodness of probability forecasts, but they have usually focused on a single aspect of the question. For example, Epstein (1962) considers the use of forecasts from a decision theoretic point of view, while Murphy and Epstein (1967) and Murphy (1972) consider only scoring rules. In this paper, we develop a general framework for comparing probabilistic forecasts and we explore the ways in which several existing methods of comparison fit into the general framework. We also hope to show how the general framework sheds new light on the overall problem of comparing forecasters.

2. Existing methods for comparing forecasters. Suppose each probability assessment must be a number from the set Ξ, a subset of $[0,1]$. For each forecaster and for each possible forecast x, consider the set of all events whose probabilities are assessed as x, and let $r(x)$ denote the long-run frequency of occurrence of these events. The function $r(x)$ is often called the forecaster's (*empirical*) *calibration curve*. If x has never been given as a forecast or if $r(x) = x$, the forecaster is said to be (*empirically*) *calibrated at* x. If the forecaster is (empirically) calibrated at every x, he is called (*empirically*) *well-calibrated*. [More general definitions of calibration are used by Dawid (1982, 1985), but we will not consider them here.] It has been proven by Dawid (1982) that in the presence of feedback, a Bayesian forecaster *expects* to be well-calibrated in an infinite sequence of forecasts. But this is relative to the forecaster's (subjective) probability distribution over the future and supposes that the forecaster has already decided how to make use of all information that will ever be learned.

It is well-known that a forecaster can be well-calibrated in a sequence of experiments if he/she always forecasts the same value x, as long as x is the "long-run" frequency of occurrence of the events of interest. Clearly, a comparison of forecasters must rest on something other than a simple comparison of calibration curves. DeGroot and Fienberg (1982a) offer a criterion which they call *refinement* for comparing well-calibrated forecasters. Refinement is related to

the concept of *sufficiency*, which we discuss in detail in Section 6. Unfortunately, since these are long-run concepts, problems of comparison in the short run cannot rest on calibration or refinement as currently understood.

To avoid the problems inherent in calibration, a method of comparing forecasters which does not rely on infinite sequences is needed. One method is the adaptation of *proper scoring rules* which Savage (1971) suggested for eliciting subjective probabilities.

DEFINITION 2.1. Consider a situation in which a forecaster must give a probability forecast for an event E. Let $Y = 1$ if E occurs and 0 if not. Let the forecaster's subjective probability of E be p. A *scoring rule* is a pair of functions $g_1(x)$ and $g_0(x)$ such that if the forecast is x and $Y = i$, the forecaster loses $g_i(x)$. A scoring rule (g_1, g_0) is *proper* if, for all p, the function

$$m(x; p) = pg_1(x) + (1 - p)g_0(x), \tag{2.1}$$

considered as a function of x for fixed p is defined, minimized and less than $+\infty$ at $x = p$. [We understand 0 times ∞ to be 0 in either summand on the right-hand side of (2.1). Also, if the right-hand side of (2.1) is $\infty - \infty$ for some x, we understand $m(x; p)$ to be $\lim_{t \to x} m(t; p)$, which we require to exist for such values of x.] A proper scoring rule is *strictly proper* if, for all p, $m(x; p)$ has its unique minimum at $x = p$.

It is easy to see that when the forecaster believes that the probability of E is p, (2.1) is the forecaster's expected loss if he/she forecasts x. Hence, a scoring rule (considered as a loss function) is proper if and only if the optimal decision is the probability of E. An old and much studied strictly proper scoring rule is the Brier score [Brier (1950)] $g_1(x) = (1 - x)^2$ and $g_0(x) = x^2$. Some authors define scoring rules as gains to the forecaster rather than losses. It will prove convenient in this paper to consider them as losses in order to make the connection with loss functions in two-decision problems more straightforward. One can use scoring rules to compare forecasters by claiming that the forecaster who performed better is the one with the smaller score after a finite number of forecasts.

3. A general method for comparing forecasters. The main goal of this section is to introduce a general framework in which the existing methods for comparing forecasters, described in Section 2, can be better understood and in which new methods with desirable properties can be introduced. The framework we choose will be that of simple two-decision problems.

Define a *simple two-decision problem* to be a problem in which there are two states of nature $Y = 0$ or 1 (unknown at the time the decision must be made) and two possible decisions d_0 and d_1, such that the loss $L(d_i, Y)$ of making decision d_i has the properties

$$L(d_1, 1) \leq L(d_0, 1),$$
$$L(d_1, 0) \geq L(d_0, 0).$$

It is easy to see that such a problem is equivalent (in terms of which decisions will be optimal) to a problem in which $L(d_i, i) = 0$ for $i = 0, 1$, $L(d_1, 0)$ and

$L(d_0, 1)$ are both nonnegative and $L(d_1, 0) + L(d_0, 1) = 1$. Such a problem can be characterized by the number $q = L(d_1, 0)$, with $0 \leq q \leq 1$, and will be called problem q.

Imagine that a decisionmaker will use the forecast x as if it were the probability that $Y = 1$ in problem q for some q. It is easy to see that the optimal decision is d_1 if $x > q$, d_0 if $x < q$ and undetermined if $x = q$. For definiteness, say that the optimal decision is d_0 if $x = q$. The loss incurred by the decisionmaker is

$$k(x; q, Y) = qI(x > q, Y = 0) + (1 - q)I(x \leq q, Y = 1), \tag{3.1}$$

where $I(B)$ denotes the indicator of the event B. Each time a forecast x_i is used, the decisionmaker in problem q would incur a loss equal to $k(x_i; q, Y_i)$. After n forecasts, let $\mathbf{x}$ be the vector of forecasts and $\mathbf{Y}$ be the vector of indicators for the events. The average loss for the decisionmaker in problem q is

$$\begin{aligned} f_{\mathbf{x}}(q, \mathbf{Y}) &= \frac{1}{n}\sum_i k(x_i; q, Y_i) \\ &= q\frac{\#(x_i > q, Y_i = 0)}{n} + (1 - q)\frac{\#(x_i \leq q, Y_i = 1)}{n}. \end{aligned} \tag{3.2}$$

Note that $k(a; q, Y) \leq k(b; q, Y)$ for all q if a is closer to Y than b is. This fact suggests the following partial ordering of forecasters.

DEFINITION 3.1. Let A and B be two forecasters and let C be some set of events. Let $\mathbf{Y}$ stand for the vector of indicators of the events in C. Let $f_A(q, \mathbf{Y})$ be the expression in (3.2) calculated from the forecasts of A and let $f_B(q, \mathbf{Y})$ be calculated from the forecasts of B. Then we say *A has performed at least as well as B* (for the events in C) if $f_A(q, \mathbf{Y}) \leq f_B(q, \mathbf{Y})$ for all q. If, in addition, $f_A(q, \mathbf{Y}) < f_B(q, \mathbf{Y})$ for some q, we say *A has performed strictly better than B.*

Clearly, not every two forecasters will be comparable according to Definition 3.1. However, we will require that any attempt to provide a more complete ordering of forecasters must agree with Definition 3.1, when that definition applies. Hence, we introduce the following:

> A **general method of comparing forecasters** is to choose a functional L on the set of all functions of the form (3.2) and a subset C of all events. We say that A has performed at least as well as B according to L for the events in C if $L(f_A) \leq L(f_B)$. The method will agree with Definition 3.1 as long as $L(h) \leq L(g)$ whenever $h(q) \leq g(q)$ for all q.

The simplest functional a decisionmaker could choose would be $L(f) = f(q_0)$ for that value of q_0 which reflects the relative risks in his/her particular decision problem. If several (or all) values of q are possible relative risks, $L(f)$ could be the integral of f with respect to some measure. The measure should put more mass on those values of q which the decisionmaker feels best describe the problems he/she must deal with. In Section 4, we will see that this is essentially equivalent to calculating a proper scoring rule. In particular, for each q, $g_Y(x) \equiv$

$k(x; q, Y)$ defines a (not strictly) proper scoring rule. In Section 6, we also consider a sense in which one f function being "better" than another means that one forecaster is better calibrated than the other. A functional which is equivalent to neither scoring rules nor calibration is $L(g) = \sup_q g(q)$. This would correspond to a minimax criterion for comparison. In fact, a whole class of functionals of this type can be generated by letting $L(g)$ be any norm of the function g.

It should be noted that many of the results in the following sections remain valid (with appropriate modifications) if $f_{\mathbf{x}}(q, \mathbf{Y})$ in (3.2) is replaced by a weighted average of the functions $k(x_i; q, Y_i)$ rather than the straight average. Interpretations of such concepts as calibration, sufficiency, dominance and refinement become more cumbersome in this case, so we will deal only with the straight average in the remainder of this paper. Unless stated otherwise, we will also assume that the subset C of events is actually the set of all events of interest.

The first step in comparing forecasters A and B might be to draw the graphs of the functions f_A and f_B to see which one was higher for what values of q. A method similar to this has been suggested by Murphy (1977). Murphy's method, however, involves several plots for each forecaster and a separate set of plots for each formulation of the decision problem (p^* in his notation). We believe that the plots of the functions (3.2) are simpler to create and their relationship to proper scoring rules (see Section 4 below) makes them more useful.

4. Proper scoring rules for comparison. Consider a proper scoring rule (g_1, g_0) as defined in Definition 2.1. In this section, we establish a relationship between such a scoring rule and the function k defined in (3.1). The main result will be a generalization of a theorem of Shuford, Albert and Massengill (1966), which can be restated in the present terminology as:

THEOREM 4.1 [Shuford, Albert and Massengill (1966)]. *If g_1 and g_0 are differentiable on $[0,1]$, g_1 is nonincreasing and g_0 is nondecreasing, then (g_1, g_0) is a proper scoring rule if and only if*

$$-pg_1'(p) = (1-p)g_0'(1-p) \quad \text{for all } 0 \le p \le 1,$$

where the prime denotes derivative.

This theorem leads to an integral representation of differentiable scoring rules which we quote verbatim from Shuford, Albert and Massengill (1966), page 129:

> $\cdots$ If we let $h(t)$ be any bounded, differentiable nonnegative function, defined on the unit interval, having the property that $h(1) = 0$ and $\sup_{|t|\le 1} |h(t)| < \infty$, then we can let

$$g_0(r) = \int_{[0,r)} h(t)\,dt \quad \text{and} \quad g_1(r) = \int_{[1-r,1)} \frac{t}{1-t} h(t)\,dt.$$

A similar representation was derived by Savage (1971). These representations are reformulated and generalized in Theorem 4.2 below. The general representation of Theorem A.9 in the Appendix makes precise the connection between proper scoring rules and the function k defined in (3.1).

Because we allow scoring rules to attain the value $-\infty$, any scoring rule satisfying the following four relations will be proper according to Definition 2.1:

$$\begin{aligned} \min\{g_0(x), g_1(x)\} &= -\infty \quad \text{for all } 0 < x < 1, \\ \max\{g_0(x), g_1(x)\} &< +\infty \quad \text{for all but finitely many } x, \\ g_0(0) &\le \inf_x g_0(x), \\ g_1(1) &\le \inf_x g_1(x). \end{aligned} \tag{4.1}$$

Such scoring rules are neither strictly proper nor very interesting. For this reason, we will not consider such scoring rules further in this paper. It is straightforward but tedious to show that all other proper scoring rules are finite for all $0 < x < 1$.

THEOREM 4.2. *Let (g_1, g_0) be a left continuous scoring rule satisfying*

$$g_i(x) = \lim_{t \to x} g_i(t) \quad \text{for } x = 0, 1, \qquad i = 0, 1, \tag{4.2}$$

and having both $g_1(t_1)$ and $g_0(t_0)$ finite for some $t_0, t_1 \in [0,1]$. This rule is proper if and only if there exists a measure $\lambda(dq)$ on $[0,1)$ such that

$$\begin{aligned} g_1(x) &= g_1(t_1) + \int_{[x, t_1)} (1 - q)\lambda(dq), \\ g_0(x) &= g_0(t_0) + \int_{[t_0, x)} q\lambda(dq) \end{aligned} \tag{4.3}$$

for all x. The scoring rule is strictly proper if and only if λ gives positive measure to every nondegenerate interval.

The proof of Theorem 4.2 relies on several lemmas and has been placed in the Appendix. The purpose of condition (4.2) is to avoid unnecessary problems which can arise when a scoring rule jumps to infinity at one of the endpoints. A result similar to Theorem 4.2 can be proven for right continuous scoring rules, and Lemma A.7 in the Appendix can be used to mix left and right continuous scoring rules into an arbitrary scoring rule. The important consequence of Theorem 4.2 occurs when both g_1 and g_0 are bounded below. Since we can add arbitrary constants to either part of a (strictly) proper scoring rule without changing the (strict) propriety, we can assume that $t_0 = 0$, $t_1 = 1$, $g_0(0) = 0$ and $g_1(1) = 0$. In this case, the score a forecaster is given when he/she forecasts x equals

$$g_1(x)I(Y = 1) + g_0(x)I(Y = 0) = \int_{[0,1)} k(x; q, Y)\lambda(dq),$$

where $k(x; q, Y)$ is defined in (3.1). After a number of forecasts, the average

score of a forecaster equals

$$\int_{[0,1)} f_{\mathbf{x}}(q, \mathbf{Y})\lambda(dq), \tag{4.4}$$

where $f_{\mathbf{x}}(q, \mathbf{Y})$ is defined in (3.2). This is the precise statement of the claim made in Section 3, that integrating f is essentially equivalent to calculating a proper scoring rule.

The representation equation (4.4) can be used to shed some light on the inconsistency of proper scoring rules discussed by Winkler and Murphy (1968). They point out that in cases in which two forecasters A and B each assign probabilities to several events, and then A and B are compared via proper scoring rules, it may turn out that A gets a lower score than B from one scoring rule and a higher score from another scoring rule. The example they give is of forecasts for several disjoint events, but the same phenomenon occurs with several forecasts in sequence. The reason for the reversal of order when one changes scoring rules is that neither $f_A(p) \leq f_B(p)$ nor $f_B(p) \leq f_A(p)$ for all p. When neither forecaster is at least as good as the other in the sense of Definition 3.1, then there exist two measures λ_1 and λ_2 such that λ_1 has more of its mass in the region where $f_A > f_B$ and λ_2 has more of its mass in the region where $f_B > f_A$. It will then follow that under the scoring rule derived from λ_1, B will have a smaller score than A and vice versa under the scoring rule derived from λ_2.

5. Examples of proper scoring rules. In light of Theorem 4.2, we can generate as many left continuous proper scoring rules as we can generate measures on $[0, 1)$. Some well-known examples are given below, along with some lesser known ones.

EXAMPLE 5.1. A simple measure that gives positive measure to every "nondegenerate" interval is $\lambda(dq) = 2\,dq$. With $t_0 = 0$ and $t_1 = 1$, (4.3) yields $g_0(x) = x^2$ and $g_1(x) = (1 - x)^2$, which is just the Brier score.

EXAMPLE 5.2. An unbounded measure, which also gives positive measure to every "nondegenerate" interval is $\lambda(dq) = dq/q(1 - q)$. Once again, with $t_0 = 0$ and $t_1 = 1$, (4.3) yields $g_0(x) = -\log_e(1 - x)$ and $g_1(x) = -\log_e(x)$, which is commonly called the logarithmic scoring rule.

EXAMPLE 5.3. A proper scoring rule which is not strictly proper is

$$g_0(x) = \begin{cases} \frac{9}{16} & \text{if } x > \frac{3}{4}, \\ x^2 & \text{if } \frac{1}{4} \leq x \leq \frac{3}{4}, \\ \frac{1}{16} & \text{if } x < \frac{1}{4}, \end{cases}$$

$$g_1(x) = \begin{cases} \frac{1}{16} & \text{if } x > \frac{3}{4}, \\ (1 - x)^2 & \text{if } \frac{1}{4} \leq x \leq \frac{3}{4}, \\ \frac{9}{16} & \text{if } x < \frac{1}{4}. \end{cases}$$

This is obtained from (4.3) by using $\lambda(dp) = 2\,dp$ for p between $\frac{1}{4}$ and $\frac{3}{4}$ and 0 outside and by setting $g_0(0) = g_1(1) = \frac{1}{16}$.

Discontinuous scoring rules can be generated by adding point masses to the measure λ which generates any continuous scoring rule. The effect of such an addition is to add a penalty to the score whenever the forecast is on the wrong side of the point where the mass is. For example, the left continuous scoring rule generated by the measure $\lambda(dq) = 2\,dq$ plus a unit point mass at $q = \frac{1}{2}$ would be Brier score plus a penalty of 1 if $x \leq \frac{1}{2}$ and $Y = 1$ or if $x > \frac{1}{2}$ and $Y = 0$.

6. How calibration fits into the general method. Throughout this section, assume that all forecasters have made forecasts for the same n events. The first result which allows us to see how calibration fits into the comparison framework described in the previous sections is easily proved by comparing $f_{\mathbf{x}}(q, \mathbf{Y})$ to the limit of $f_{\mathbf{x}}$ from the left of q.

THEOREM 6.1. *A forecaster is empirically well-calibrated if and only if his function* $f_{\mathbf{x}}(q, \mathbf{Y})$ *defined in* (3.2) *is continuous in* q.

In light of Theorem 6.1, if being calibrated were all that was of interest, one could define a sense in which forecaster A has performed better than B if f_A is closer to being continuous than f_B. On the surface, such a definition would have little or no connection with Definition 3.1, since f_A can be continuous, f_B discontinuous and $f_A(q, \mathbf{Y}) \geq f_B(q, \mathbf{Y})$ for all q. For example, suppose there were only two events being forecast, $Y_1 = 0$ and $Y_2 = 1$. Suppose A forecasts 0.5 both times, while B forecasts 0.2 for the first event and 0.8 for the second. Then A is calibrated while B is not. But $f_A(q, \mathbf{Y}) \geq f_B(q, \mathbf{Y})$ for all q with strict inequality for q between 0.2 and 0.8. It is clear after the fact that B's forecasts were better than A's.

There is, however, a sense in which, for each forecaster B who is not calibrated, there "exists" a forecaster A who is calibrated and has performed better than B.

DEFINITION 6.2. Consider a single forecaster B. For each forecast x given by B, let

$$r_B(x) = \frac{\#(x_i = x, Y_i = 1)}{\#(x_i = x)} \tag{6.1}$$

be the *empirical calibration curve* for forecaster B. Let A be a mythical forecaster who gave forecast $r_B(x)$ each time B gave forecast x, for all x. We say that forecaster A is the (*empirically*) *calibrated version* of forecaster B.

Note that forecaster A in Definition 6.2 is automatically well-calibrated. We say that A is a mythical forecaster, because he would have to know the values of $r_B(x)$ *before* seeing the Y_i in order to make the forecasts he does. Clearly, if B is already well-calibrated, then B and A give the same forecasts. If the set of

possible forecasts Ξ is not the entire unit interval, then A gives invalid forecasts whenever $r_B(x) \notin \Xi$. Without going into too much detail, one could define D to be an *almost calibrated version of B relative to* Ξ by having D give the closest forecast to $r_B(x)$ which is in Ξ and is between x and $r_B(x)$ whenever B gives forecast x. Theorem 6.3 below still holds if the word "almost" is inserted before "calibrated version," but some of the other results do not hold after that change. We will not consider almost calibrated versions any further in this paper.

THEOREM 6.3. *Let B be a forecaster and A his calibrated version. If B is not well-calibrated, then A has performed strictly better than B in the sense of Definition* 3.1.

PROOF. Using formula (3.2), we can write, after some simplification,

$$\begin{aligned} & f_B(p) - f_A(p) \\ & \quad = p\#(x_i > p,\ r_B(x_i) \le p) \\ & \quad\quad - \#(x_i > p,\ r_B(x_i) \le p,\ Y_i = 1) \\ & \quad\quad + \#(x_i \le p,\ r_B(x_i) > p,\ Y_i = 1) - p\#(x_i \le p,\ r_B(x_i) > p). \end{aligned} \tag{6.2}$$

Consider the two terms in the first row of (6.2). Since we are counting trials for which $r_B(x_i)$ is less than p, the proportion of such trials for which $Y_i = 1$ is clearly less than p by (6.1). Hence the first row of (6.2) is nonnegative. Similarly, the second row of (6.2) is nonnegative and $f_B(p) \ge f_A(p)$ for all p. If B is not well-calibrated, then for each p which is between x_i and $r_B(x_i)$ for some i, one of the rows of (6.2) will be strictly positive. □

The relationship between a forecaster and his calibrated version is a special case of a more general relationship called *sufficiency* by DeGroot and Fienberg (1982a).

DEFINITION 6.4. A function $h(x|y)$ defined for all x and y in a finite set of possible forecasts is called a *stochastic transformation* if

$$h(x|y) \ge 0 \quad \text{for all } x \text{ and } y,$$

$$\sum_x h(x|y) = 1 \quad \text{for all } y.$$

DEFINITION 6.5. For a forecaster A, let $n_A(x) = \#(x_i = x)/n$ and let $r_A(x)$ be forecaster A's calibration curve. Similarly, define n_B and r_B for another forecaster B. We say that A is (*empirically*) *sufficient* for B if there exists a stochastic transformation $h(x|y)$, defined for all x in the set of forecasts given by B and all y in the set of forecasts given by A, such that

$$\sum_y h(x|y) n_A(y) = n_B(x) \qquad \text{for all } x,$$

$$\sum_y h(x|y) r_A(y) n_A(y) = r_B(x) n_B(x) \quad \text{for all } x.$$

The word *empirically* is in parentheses in Definition 6.5 because the context of this definition is slightly different from that of DeGroot and Fienberg (1982a). DeGroot and Fienberg are dealing with probabilities, not frequencies from finite samples. However, since frequencies in finite samples behave like probabilities, many of the theorems of DeGroot and Fienberg can be used in the present context without reproving them. We will make free use of this fact in the ensuing discussion. We will discuss the distinction between probabilities and frequencies again in Section 10.

A sense in which being empirically sufficient means being better is given by the next theorem.

THEOREM 6.6. *A forecaster A is empirically sufficient for B if and only if the calibrated version of A has performed at least as well as the calibrated version of B in the sense of Definition* 3.1.

PROOF. It can easily be shown that a forecaster and his calibrated version are empirically sufficient for each other and that empirical sufficiency is transitive. Hence, it suffices to consider forecasters A and B who are well-calibrated. Hence, we may assume $r_A(x) = r_B(x) = x$ for all x. In this case, we can write (after some simplification)

$$f_B(p) - f_A(p) = \sum_{x \le p} (p - x)\{n_A(x) - n_B(x)\},$$

where the summation is over all forecasts (given by either forecaster) that are less than or equal to p. Theorem 3 of DeGroot and Fienberg (1982a), when translated to handle the empirical case, applies to finish the proof. □

Consider, once again, the simple two-decision problems discussed in Section 3. Theorem 6.6 says that a forecaster A is empirically sufficient for B if and only if we would have done no worse to use the calibrated forecasts of A than the calibrated forecasts of B no matter what problem q we were making decisions in. Note that if A is very poorly calibrated, it is possible for B to have performed better than A, while A was empirically sufficient for B. This would be the case, for example, if A always gave one of the forecasts 0 and 1, but was always wrong. The relationship between sufficiency of forecasters and sufficiency in the theory of comparison of experiments will be taken up in Section 10.

As a final note, the results of this section are particularly uninteresting in the case in which B gives each forecast only once. In this case, the calibrated version of B is the perfect forecaster, the one who says only 0 or 1 and is always correct. Such considerations might lead to more restrictive definitions of calibration such as the ones in Dawid (1982, 1985). Dawid's definitions, however, have less intuitive appeal than the more standard definition given above and lead to complications which have not yet been resolved [see Schervish (1985b)].

7. Dominance considerations. Recently, Vardeman and Meeden (1983) introduced three partial order relations among forecasters. We translate these relations into the notation introduced above.

DEFINITION 7.1. Forecaster A (*empirically*) *rain dominates* forecaster B if, for all q,

$$\sum_{x \le q} r_A(x) n_A(x) \le \sum_{x \le q} r_B(x) n_B(x). \tag{7.1}$$

Forecaster A (*empirically*) *dry dominates* forecaster B if, for all q,

$$\sum_{x \le q} \{1 - r_A(x)\} n_A(x) \ge \sum_{x \le q} \{1 - r_B(x)\} n_B(x). \tag{7.2}$$

Forecaster A (*empirically*) *dominates* B if both (7.1) and (7.2) hold for all q.

The authors then proceed to prove several mathematical properties related to dominance. For example, if A and B are both well-calibrated and A either rain or dry dominates B, then A is sufficient for B.

In this section, we will explore how these dominance concepts fit into the general method of comparing forecasters. It is easy to see, for example, that A empirically dry (rain) dominates B if and only if the first (second) term in (3.2) for f_A is no larger than the corresponding term for f_B for all q. Hence, A dry dominates B if and only if A performs at least as well as B on dry days, that is, the set C is the set of all trials on which it does not rain. Similarly, A rain dominates B if and only if A performs at least as well as B on rainy days. There is also a connection between rain/dry dominance and overall performance. The following theorem is easy to prove, so the proof will be omitted. Note that there is no calibration condition for either A or B.

THEOREM 7.2. *If A dominates B, then A performs at least as well as B. If A dry (rain) dominates B and B performs at least as well as A, then B rain (dry) dominates A.*

In a sense, Theorem 7.2 says that either rain or dry dominating another forecaster is a hedge against having the other forecaster perform better, but is no guarantee. For example, if A dry dominates B, then the only way B could perform at least as well as A is for B to rain dominate A to at least as great an extent as A dry dominates B. One can easily construct examples of all of the cases not covered by Theorem 7.2. For example, one could have A rain dominate B and B dry dominate A, while either $f_A \le f_B$ or $f_A \ge f_B$ or neither. Similarly, one could have A rain dominate B and B not dry dominate A, while either $f_A(q, \mathbf{Y}) \le f_B(q, \mathbf{Y})$ for all q or not. Also, we could have A perform at least as well as B, but A neither rain nor dry dominates B.

8. Proper scoring rules and calibration. Theorem 6.3 states that the calibrated version A of an ill-calibrated forecaster B has performed strictly better than B. Hence, the average score on any proper scoring rule should be smaller for A than for B. The purpose of this section is to consider exactly how much smaller the average score is. For simplicity consider only scoring rules (g_0, g_1) which are left continuous and which satisfy $g_0(0) = 0 = g_1(1)$. Then there exists a measure λ on the unit interval such that the average score of

forecaster A is

$$\int_{[0,1)} f_A(q, \mathbf{Y})\lambda(dq)$$

and the average score of forecaster B is

$$\int_{[0,1)} f_B(q, \mathbf{Y})\lambda(dq).$$

We can write the average score of B as

$$\sum_x n_B(x)\{r_B(x)g_0(x) + [1 - r_B(x)]g_1(x)\}, \tag{8.1}$$

where the sum is over all of the distinct forecasts which B gives. DeGroot and Fienberg [(1982b), Theorem 4] write (8.1) as $S_1 + S_2$, where S_2 is just the average score of forecaster A, the calibrated version of B and S_1 is the excess, which they write as

$$S_1 = \sum_x n_B(x)\{r_B(x)(g_1(x) - g_1[r_B(x)]) + [1 - r_B(x)][g_0(x) - g_0(r_B(x))]\}.$$

It is not easy to see how this expression measures the degree to which B is not calibrated, except that it is zero when B is calibrated and positive otherwise. However, if we use the integral representation of the average score, we see that

$$S_1 = \sum_x \int_{[x,\, r_B(x))} n_B(x)\{r_B(x) - q\}\lambda(dq), \tag{8.2}$$

where the range of integration is to be understood in the sense of Definition A.8 (in the Appendix) if $x > r_B(x)$. The integrand in (8.2) is nonnegative over the range of integration and the closer x is to $r_B(x)$, the shorter that range is. We can also see that it is particularly bad to be ill-calibrated in regions with high measure under λ. Recall that λ gives high measure to regions which include q values corresponding to those simple two-decision problems which are most important. Poor calibration in such a region would correspond to the forecast x being too often on the wrong side of the cutoff q in problem q, leading the decisionmaker to make the more costly decision too often.

Just as S_1 has an interpretation as a measure of how far forecaster B is from being calibrated, S_2, which is the accumulated score of the calibrated version of B, has an interpretation as a measure of how good B was at distinguishing trials. If $r_B(x)$ is very different from $r_B(y)$ when x and y are different, then B was good at distinguishing trials, even if he distinguished them incorrectly. For example, the forecaster who always forecasts 0 or 1 and is always wrong distinguishes trials as well as the perfect forecaster who is always right. DeGroot and Fienberg (1982b) say that forecaster A is *at least as refined as* forecaster B if A is at least as good as B at distinguishing trials in some sense. The exact sense will not concern us, since whenever A and B are well-calibrated, DeGroot and Fienberg prove that A is at least as refined as B if and only if A is sufficient for B. To see how S_2 can be interpreted as a measure of how refined (or how

"sufficient") B is, write S_2 as

$$S_2 = \sum_x n_B(x)\psi[r_B(x)],$$

where

$$\psi(t) = tg_1(t) + (1 - t)g_0(t).$$

The function ψ is concave, since it is the Bayes risk in the decision problem with parameter $Y \in \{0,1\}$, decisions x in the interval $[0,1]$ and loss function $L(x, Y) = g_Y(x)$. A concave function on a bounded interval is smallest near at least one of the ends of the interval. So S_2 is small if $r_B(x)$ is always near 0 or near 1. That is, B is refined in so much as his forecasts distinguish groups of trials with far apart values of r_B.

In summary, a forecaster accumulates a high score for two reasons:

1. He is not well-calibrated.
2. His calibrated version is not perfectly refined.

This result can be easily misinterpreted. It would seem to suggest that one should strive to be well-calibrated, since that will reduce one's function $f(p)$. However, attempts to become more calibrated may make one less refined. Unfortunately, unless one is failing to learn from experience, the only way to become better calibrated is to figure out what is going to happen before it does. Hence, one should strive to figure out what will happen before it happens. A more reasonable interpretation of these results is given in Section 10.

9. An example of the use of the general method. During the summer of 1982, the author kept track (in a somewhat nonsystematic manner) of the forecasts for rain on the current day and for the next day. The variable Y was taken to be 1 if the National Oceanic and Atmospheric Administration recorded at least 0.01 in. of rain for the given day and 0 otherwise. There were 47 days for which the author had two forecasts, the one given on that day and the one given the day before. We will call the forecaster who forecasts today's rain today A and the one who forecasts today's rain yesterday B. The functions n and r for each forecaster are given in Table 1. One can easily see that neither set of forecasts is exceptionally well-calibrated. The functions $f_A(q, \mathbf{Y})$ and $f_B(q, \mathbf{Y})$ are plotted in Figures 1 and 2 along with the corresponding functions for their calibrated versions. Figure 1 shows the forecasters compared to each other (I) and compares their calibrated versions to each other (II). The calibrated version of A has performed at least as well as the calibrated version of B for all q, indicating that A is more refined than B. Surprisingly, however, there is a substantial range of values of q, from 0.3 to 0.5, over which B performed better than A. Figure 2 shows each forecaster compared to his calibrated version. [Today's forecast given today is (I) and today's forecast given yesterday is (II)]. In this plot, one can see that B is more nearly calibrated than A.

To compare the two forecasters, one might wish to use several proper scoring rules. Since neither forecaster has performed better than the other, different

TABLE 1
Values of $n(x)$ and $r(x)$ for current and previous day forecasts. A denotes today's forecast given today. B denotes today's forecast given yesterday.

x	$n_A(x)$	$r_A(x)$	$n_B(x)$	$r_B(x)$
0.0	16	0.063	16	0.063
0.2	3	0.000	0	
0.3	10	0.400	6	0.167
0.4	3	0.000	8	0.250
0.5	2	0.500	4	0.750
0.6	3	0.667	3	0.333
0.7	3	0.667	7	0.714
0.8	1	0.000	1	1.000
0.9	0		1	1.000
1.0	6	1.000	1	1.000

scoring rules will rank them differently. For example, the average Brier scores are 0.1402 and 0.1485 for A and B, respectively, while the average scores for the rule of Example 5.3 are 0.0961 and 0.0761, respectively. This is due to the fact that B was better than A for forecasts in the middle of the range, while A was better than B to an even greater extent for extreme forecasts.

10. Beliefs about forecasters. In the previous sections, we have dealt almost exclusively with a finite set of trials in which one or more forecasters gave a probability for the event $E_i = \{Y_i = 1\}$, and we learned whether or not E_i occurred. We judged forecasters based on how close the forecasts were to the observed values of Y_i. What if we are interested in expressing our beliefs about a future prediction? We cannot judge a future forecast about Y based on how close it is to Y until after Y is observed. If the forecast which will be given is X, then we must consider the joint distribution of (X, Y). By "distribution" here, we mean the subjective probability distribution that some decisionmaker has over the forecasts and the forecasted events. In a manner similar to (3.2), write

$$\phi_X(q) = q \Pr(X > q, Y = 0) + (1 - q)\Pr(X \le q, Y = 1).$$

The results of the previous sections which concerned the accumulated score can be easily translated to results concerning the expected score from a proper scoring rule by replacing $f_X(q, Y)$ by $\phi_X(q)$ and making some minor terminological changes.

We will call a forecaster *probability calibrated* [following Lindley (1982)] if

$$\rho(x) = \Pr(Y = 1 | X = x) = x$$

for all possible forecasts x. The function $\rho(x)$ can be called the forecaster's *probability calibration curve* in analogy to the function $r(x)$. Similarly define $\nu(x)$ to be $\Pr(X = x)$ in analogy to $n(x)$. Call A the *probability calibrated version of* B if A gives forecast $\rho(x)$ after B gives forecast x. Then Theorem 6.3 translates to say that $\phi_A(q) \le \phi_B(q)$ for all q. The conclusion that one should

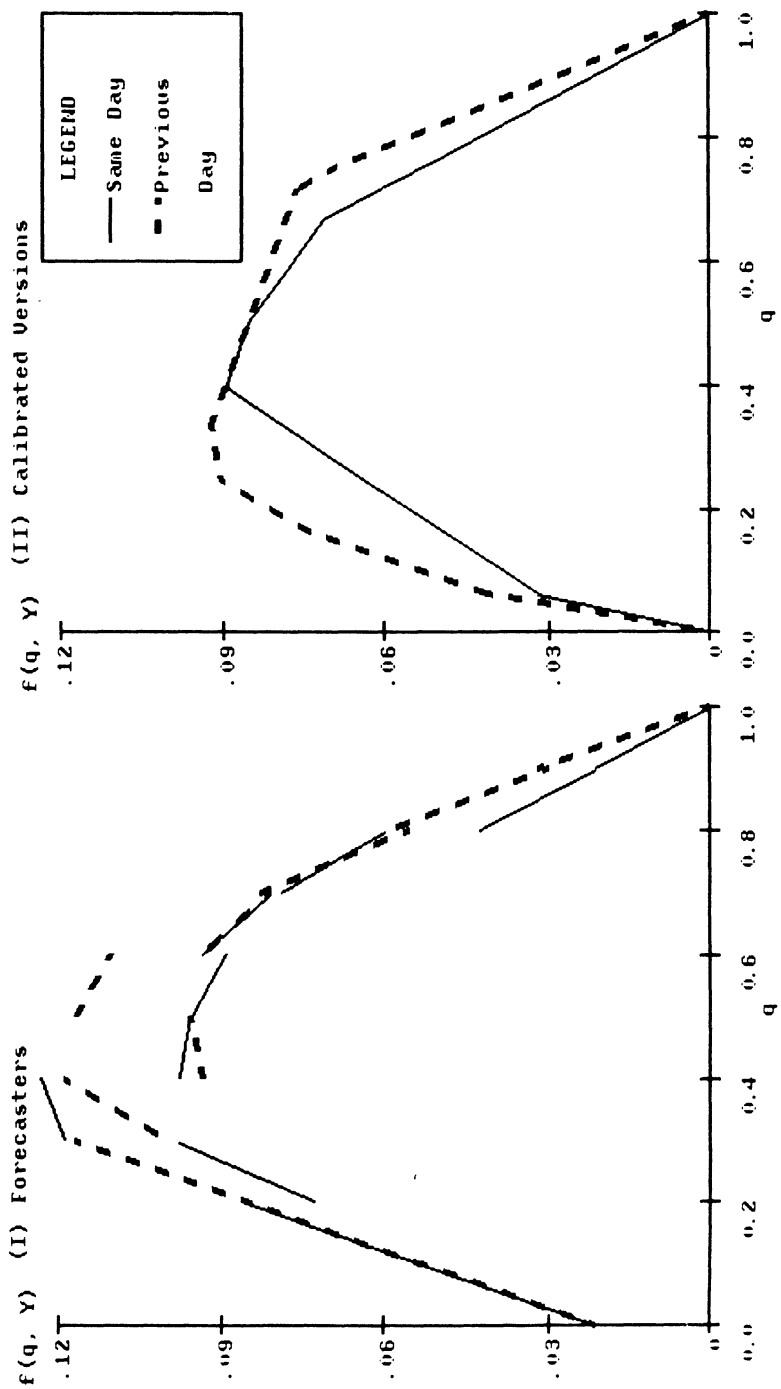

FIG. 1. *Comparing two forecasters to each other.*

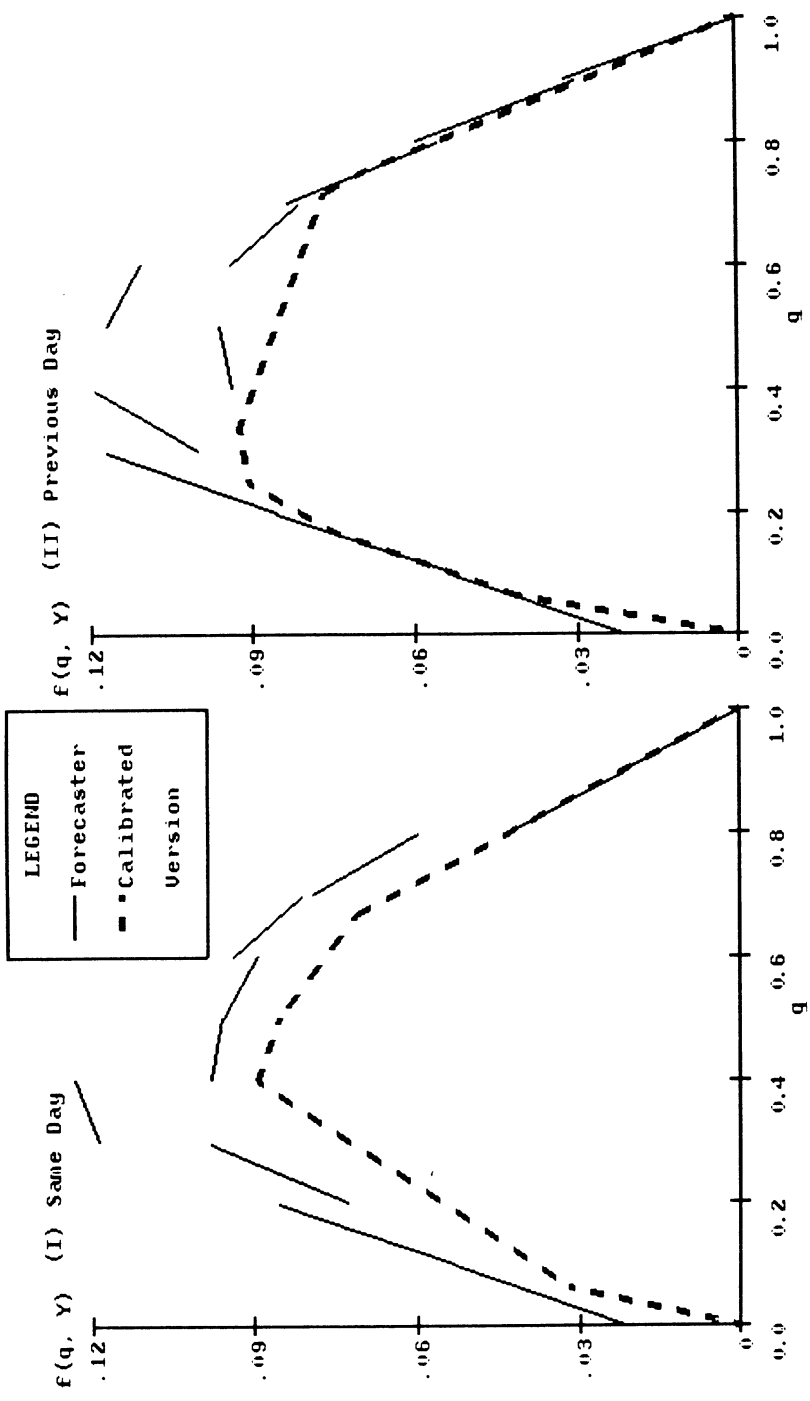

FIG. 2. *Comparing two forecasters to their calibrated versions.*

strive to be calibrated now translates to the reasonable, and almost forgone, conclusion that if one is going to use a forecast x as if it were $\Pr(Y = 1)$, one would do best if $x = \Pr(Y = 1)$.

The concept of probability sufficiency is complicated by the fact that the set of possible forecasts may not be finite, whereas in the empirical case, the set of observed forecasts is always finite. We take the approach of Blackwell (1951) in the following definition.

DEFINITION 10.1. Consider two forecasters A and B. We say that A is (*probability*) *sufficient* for B if there exists a function $H(E|y)$ with the following properties:

(i) $H(\cdot|y)$ is a probability over $[0,1]$ for each $y \in [0,1]$.
(ii) $H(E|\cdot)$ is a Borel-measurable function for each Borel subset E of $[0,1]$.
(iii) $\mu_{iB}(E) = \int H(E|y)\mu_{iA}(dp)$ for $i = 0, 1$,

where, for any forecaster C,

$$\mu_{iC}(E) = \Pr(C\text{'s forecast} \in E | Y = i).$$

The function H takes the place of the stochastic transformation h in Definition 6.5 because it is not necessary for the set of possible forecasts to be finite. For the case in which the set of possible forecasts is finite, a simple translation of Theorem 6.6 to the case of probabilities says that A is probability sufficient for B if and only if the expected loss from using A's calibrated forecasts is no greater than the expected loss from using B's forecasts (calibrated or not) regardless of which problem q we are making decisions in. For the more general case, we state the following theorem, which can be proved by applying Theorems 5, 8 and 10 of Blackwell (1951) and Theorem 8.3.2 of Blackwell and Girshick (1954).

THEOREM 10.2. *If A and B are probability calibrated forecasters, then A is probability sufficient for B if and only if $\phi_A(q) \le \phi_B(q)$ for all q.*

This, then, makes the connection with sufficiency in comparison of experiments more clear. Forecaster A is probability sufficient for forecaster B if the experiment which provides us with A's calibrated forecast is sufficient for the experiment which provides us with B's (calibrated) forecast.

Finally, consider the probabilistic analogs of the dominance criteria of Vardeman and Meeden (1983). We might say that A *probability rain dominates* B if, for all q,

$$\Pr(X_A \le q, Y = 1) \le \Pr(X_B \le q, Y = 1)$$

and A *probability dry dominates* B if, for all q,

$$\Pr(X_A \le q, Y = 0) \ge \Pr(X_B \le q, Y = 0).$$

A simple analog of Theorem 7.2 applies and will not be stated here. Unfortunately, there is a serious drawback to these criteria. Even if one knew that A dry

dominated B, one would be hard pressed to make use of the information, because all it means is that the expected loss from using A's forecast is smaller than that from using B's on dry days. In order to use the information, one must either know whether or not it is a dry day, in which case one no longer needs the forecasts, or one must decide how likely it is that the day is dry, which begs the original question the forecasters are trying to answer.

11. Discussion. In this paper, we have introduced a general method for comparing probability assessors which includes, as special cases, several other popular methods. Studying the properties of this general method has also aided our understanding of the problem of comparing forecasters in two important ways. First, it has highlighted some of the strengths and weaknesses of existing methods. For example, we saw that scoring rules are just a way of averaging all simple two-decision problems into a single, more complicated, decision problem. We also saw how calibration by itself is not a measure of how good a forecaster is. Second, the study of the general method has helped to clarify the difference between comparing forecasters based solely on their past performance and comparing them based on expected performance in the future. In particular, we saw how it was easy to use the concept of empirical dominance to compare two forecasters after seeing their forecasts, but it is not so easy to use probability dominance when comparing future forecasts.

The distinction between these two problems (comparison after the fact and comparison of future forecasts) has caused a good deal of confusion in the literature [see Dawid (1982), Sections 6 and 7, and Dawid (1985) for examples]. One cannot evaluate a forecaster *today* based on how good his forecasts *will be* in the future. One can evaluate him today based on how good *one believes* his forecasts will be in the future and/or based on how good his forecasts *were* in the past. The confusion so prominent in discussions about calibration, for example, is caused by an attempt to use results designed for evaluations based on past performance as if they were valid for evaluations based on beliefs about the future. The "past performance" used is usually the infinite future. For example, the idea that one should strive to be empirically calibrated is a misinterpretation of the result that the probability calibrated version of a forecaster is better (to the decisionmaker) than the forecaster himself. Lindley, Tversky and Brown (1979) and Lindley (1982) present methods for improving a forecaster who is not probability calibrated. On the other hand, there is no way to take a forecaster who is not empirically calibrated and redo his forecasts before they are known to make him empirically calibrated. This has been proven by Oakes (1985) and Schervish (1985a).

In light of the above problems, the goal of this paper has been to consider only those evaluation procedures which can be performed and which have meaningful interpretations at the time they are performed. In Section 3, we introduced the framework for a general method of evaluating and comparing forecasters based on past performance. In Section 10, we translated the results valid in that framework to results useful for evaluations based on beliefs about future forecasts. Although some of the results pertaining to comparison based on past

observations have the same forms as those pertaining to comparison based on beliefs, both their interpretations and their mechanics are very different. For example, consider the forecasters of Section 9. Just because B performed better than A for values of q near 0.4, I would not be more inclined to follow yesterday's forecast for today's weather than today's (even if my q were near 0.4) because of my high prior probability that today's forecast is better. However, if I have both forecasts available, I could try to combine them into a single, hopefully better, forecast. The problem of combining forecasts, however, is a separate problem, which also deserves serious attention.

APPENDIX

In this section, we will prove that g_1 and g_0 must be monotone and remove the restriction that they be continuous. The present treatment begins with a lemma.

LEMMA A.1. *Let (g_1, g_0) be a (strictly) proper scoring rule, possibly attaining infinite values on the closed interval $[0, 1]$. Then $g_1(x)$ is (strictly) decreasing in x and $g_0(x)$ is (strictly) increasing in x.*

PROOF. Let $0 < b - a < 1$. We have, from Definition 2.1,

$$\begin{aligned} m(a; b) &\geq m(b; b), \\ m(b; a) &\geq m(a; a), \end{aligned} \tag{A.1}$$

with strict inequality in both in the strictly proper case. It is straightforward to show that, for proper scoring rules which violate (4.1), the restrictions on m in Definition 2.1 imply that the only possible infinite values are $g_1(0) = \infty$, $g_1(1) = -\infty$, $g_0(0) = -\infty$ or $g_0(1) = \infty$. It is clear from (A.1) and (2.1) that $g_1(x) \geq g_1(1)$ and $g_0(1) \geq g_0(x)$ for all $x < 1$ and that $g_0(x) \geq g_0(0)$ and $g_1(0) \geq g_1(x)$ for all $x > 0$ (with strict inequalities in the strictly proper case). So, it suffices to consider $0 < a < b < 1$. In this case, the quantities $c = g_1(a) - g_1(b)$ and $d = g_0(a) - g_0(b)$ are both finite. It follows that the inequalities in (A.1) can be written

$$bc + (1 - b)d \geq 0 \geq ac + (1 - a)d, \tag{A.2}$$

with strict inequalities in the strictly proper case. Subtract $ac + (1 - b)d$ from all three parts of (A.2) and divide by $b - a > 0$ to obtain

$$c \geq -\frac{ac + (1 - b)d}{b - a} \geq d. \tag{A.3}$$

Since (A.3) implies $c \geq d$, add $(1 - b)(c - d) \geq 0$ to the left of (A.2) and $a(d - c) \leq 0$ to the right to obtain $c \geq 0 \geq d$ (with strict inequalities in the strictly proper case). □

Also, since proper scoring rules are monotone, they are continuous except possibly for countably many jump discontinuities. In case $m(0; q)$ or $m(1; q)$ is

ever $\infty - \infty$, define its value via the limit, which either exists or is infinite because the sum of monotone functions has bounded variation. The next lemma describes a useful continuity property of proper scoring rules.

LEMMA A.2. *The function $m(x; p)$ defined in* (2.1), *considered as a function of x for fixed p, is continuous at $x = p$ if g_1 and g_0 are both bounded in a neighborhood N of $x = p$.*

PROOF. Let p be fixed and assume $m(x; p)$ has a jump discontinuity (the only possible kind of discontinuity for the difference of two monotone functions) such that

$$\lim_{x \downarrow p} m(x; p) = c + m(p; p)$$

with $c > 0$. [This can only happen if $p < 1$. If $p = 1$, the limit from the left must be greater than $m(p; p)$. The other cases can be handled in a fashion similar to the treatment below.] Let δ be small enough so that the sum of the absolute values of all of the other jump discontinuities of $m(x; p)$ for $|x - p| \leq \delta$ is at most $c/4$ and so that the continuous part of $m(x; p)$ varies by at most $c/4$ for $|x - p| \leq \delta$. Because $m(x; y)$ is linear in y for each x and because g_1 and g_0 are both bounded for $x \in N$, it follows that there exists ϵ such that $|m(x; p^*) - m(x; p)| < c/4$ for all $x \in N$ and $|p - p^*| < \epsilon$. Now choose any $p^* > p$ such that $|p - p^*| < \epsilon$, $p^* \in N$ and $|p - p^*| < \delta$. The following contradictory string of inequalities obtains:

$$\begin{aligned} m(p^*; p^*) &\leq m(p; p^*) < m(p; p) + c/4 \\ &< m(p^*; p) - c/4 < m(p^*, p^*). \end{aligned}$$

□

The following corollary to Lemma A.2 says that the discontinuities of g_1 and of g_0 are intimately tied together.

COROLLARY A.3. *For all p strictly between* 0 *and* 1,

$$(1 - p)\left[g_0(p) - \lim_{x \downarrow p} g_0(x)\right] = -p\left[g_1(p) - \lim_{x \downarrow p} g_1(x)\right]$$

and

$$(1 - p)\left[g_0(p) - \lim_{x \uparrow p} g_0(x)\right] = -p\left[g_1(p) - \lim_{x \uparrow p} g_1(x)\right].$$

The possible discontinuities of a proper scoring rule at 0 and at 1 are of a slightly different nature than the others. The following three lemmas make this more explicit. Their proofs are omitted because discontinuities at the endpoints are not of great interest, especially in light of Lemma A.6.

LEMMA A.4. *Suppose (g_1, g_0) is a (strictly) proper scoring rule. Let $c = \lim_{t \downarrow 0} g_1(t)$ and $d = \lim_{t \uparrow 1} g_0(t)$. Then any scoring rule (h_1, h_0) which equals (g_1, g_0) for all x strictly between* 0 *and* 1 *is also (strictly) proper so long as it satisfies $h_1(0) \geq g_1(0)$ and $h_0(1) \geq g_0(1)$.*

LEMMA A.5. *Suppose* (g_1, g_0) *is a proper scoring rule. If* $g_0(0) < \lim_{t \downarrow 0} g_0(t)$, *then* $g_1(0) = \infty$. *Similarly, if* $g_1(1) < \lim_{t \uparrow 1} g_1(t)$, *then* $g_0(1) = \infty$.

LEMMA A.6. *A scoring rule satisfying* $g_0(0) < \lim_{t \downarrow 0} g_0(t)$ *and/or* $g_1(1) < \lim_{t \uparrow 1} g_1(t)$ *is (strictly) proper if and only if the scoring rule* (h_1, h_0) *is (strictly) proper, where* $(h_1, h_0) = (g_1, g_0)$ *for all* x *strictly between* 0 *and* 1 *and, for* x *equal* 0 *or* 1, $h_i(x) = \lim_{t \to x} g_i(t)$ *for* $i = 0, 1$.

Lemma A.6 is the justification for only dealing with scoring rules which satisfy (4.2).

The next property of proper scoring rules is trivial and is stated without proof. All it says is that a proper scoring rule is a mixture of continuous, left continuous and right continuous proper scoring rules. One term in the mixture is needed for each point at which g_1 (and hence g_0) is discontinuous. Corollary A.3 guarantees that the same set of mixing coefficients works for both g_1 and g_0.

LEMMA A.7. *If* (g_1, g_0) *is a proper scoring rule, then there exists a continuous proper scoring rule* (g_{c_1}, g_{c_0}) *and (at most) countably many left continuous proper scoring rules* $(g_{L_{1i}}, g_{L_{0i}})$ *and right continuous proper scoring rules* $(g_{R_{1i}}, g_{R_{0i}})$ *such that for* $j = 0, 1$,

$$g_j = g_{cj} + \sum_{i=1}^{\infty} (g_{L_{ji}} + g_{R_{ji}}).$$

Each of the functions $g_{L_{ji}}$ *and* $g_{R_{ji}}$ *is a step function with a single jump discontinuity at a point* x_i *(the same point for* R *as for* L *and the same for* $j = 1$ *as for* $j = 0$, *with the sizes of the jumps constrained by Corollary* A.3).

In view of Lemma A.7, we will prove Theorem 4.2 for left continuous proper scoring rules and note that the proof is analogous for right continuous scoring rules. The two versions can then be combined together with Lemma A.7 for the most general result. Because Theorem 4.2 involves integrals over intervals with arbitrary endpoints in $[0, 1]$, we need the following definition first.

DEFINITION A.8. Let $\mu(dx)$ be a measure on $[0, 1]$ and let f be any measurable function. If $b < a$, define

$$\int_{[a, b)} f(x)\mu(dx) = -\int_{[b, a)} f(x)\mu(dx)$$

and define the interval $[a, b)$ to be the interval $[b, a)$. By convention, the interval $[a, a)$ is empty. This definition allows us to refer to the interval between two numbers a and b, including the smaller but excluding the larger as $[a, b)$ regardless of which is larger.

PROOF OF THEOREM 4.2. First we prove the "if" parts. Since we can add arbitrary constants to either or both parts of a proper scoring rule without

changing its propriety, we will assume that $g_1(t_1) = g_0(t_0) = 0$. We can calculate $m(x; p) - m(p; p)$ as

$$p\int_{[x, p)}(1 - q)\lambda(dq) + (1 - p)\int_{[p, x)} q\lambda(dq),$$

which equals

$$\int_{[x, p)}(p - q)\lambda(dq). \tag{A.4}$$

The integrand in (A.4) is positive over the range of integration if $x \leq p$ and is nonpositive over the range if $x \geq p$. But in the latter case, the integral must be multiplied by -1 (by Definition A.8) and the integral in (A.4) is nonnegative in either case. This proves that a scoring rule defined by (4.3) is proper. If λ gives positive measure to every nondegenerate interval, then the integral in (A.4) will be strictly positive whenever $x \neq p$ and the scoring rule will be strictly proper.

The "only if" parts require more work. We assume the scoring rule (g_1, g_0) is proper. Lemma A.1 says that g_1 is monotone decreasing and g_0 is monotone increasing. Define three measures on the interval $[0, 1)$ by defining them for all intervals $[a, b)$ with $b > a$ and extending to all Borel sets:

$$\begin{aligned} \lambda_0[a, b) &= g_0(b) - g_0(a), \\ \lambda_1[a, b) &= g_1(a) - g_1(b), \\ \lambda &= \lambda_0 + \lambda_1. \end{aligned} \tag{A.5}$$

By (4.2), we may assume that all three measures assign mass 0 to the singleton $\{0\}$. It now follows that the measures are σ-finite. Clearly λ_0 and λ_1 are both absolutely continuous with respect to λ. Let $h_0(p)$ and $h_1(p)$ be their respective Radon–Nikodyn derivatives. Since λ is the sum of λ_0 and λ_1,

$$h_0(q) + h_1(q) = 1 \quad \text{a.e. } [\lambda]. \tag{A.6}$$

It follows from (A.5) and Definition A.8 that, for all x,

$$\begin{aligned} g_0(x) &= g_0(t_0) + \int_{[t_0, x)} h_0(q)\lambda(dq), \\ g_1(x) &= g_1(t_1) + \int_{[x, t_1)} h_1(q)\lambda(dq). \end{aligned} \tag{A.7}$$

In light of (A.6), all that remains to show is that $h_0(q) = q$ a.e. $[\lambda]$. Since the scoring rule is proper, we know that $m(x; p) - m(p; p) \geq 0$ for all x and p. Using (A.6) and (A.7), we write

$$\int_{[x, p)}\{p - h_0(q)\}\lambda(dq) \geq 0 \quad \text{for all } x, p. \tag{A.8}$$

Lemma 20.54 of Hewitt and Stromberg [(1965), page 367] together with Corollary A.3 can be used to show that (A.8) implies $h_0(q) = q$ a.e. $[\lambda]$.

In the strictly proper case, if λ gave zero measure to a nondegenerate interval $[a, b)$, then for all x and p in $[a, b)$, $m(x; p) = m(p; p)$ by (A.4) and the scoring rule would not be strictly proper. □

Finally, we may wish to consider a general version of Theorem 4.2 for the case in which the scoring rule is bounded below. First, define

$$h(x; q, Y) = qI(x \geq q, Y = 0) + (1 - q)I(x < q, Y = 1) \tag{A.9}$$

as the right continuous version of $k(x; q, Y)$.

THEOREM A.9. *Assume the functions g_0 and g_1 are bounded from below and satisfy* (4.2). *Then (g_1, g_0) is a proper scoring rule if and only if there exist measures λ_i on the interval $[0,1]$ and nonnegative constants $\{\alpha_i\}$ and $\{\beta_i\}$ for $i = 0, 1, \ldots$ such that each of the measures λ_i for $i = 1, 2, \ldots$ is a unit point mass at a different point q_i, λ_0 is nonatomic and*

$$\{g_1(x) - g_1(1)\}I(Y = 1) + \{g_0(x) - g_0(0)\}I(Y = 0)$$

$$= \sum_i \int_{(0,1)} \{\alpha_i k(x; q, Y) + \beta_i h(x; q, Y)\}\lambda_i(dq).$$

The rule is strictly proper if and only if, in addition, the measure

$$\sum_i (\alpha_i + \beta_i)\lambda_i$$

assigns positive probability to every "nondegenerate" interval.

Acknowledgments. The author would like to thank Professor Abraham Wender for carefully reading an earlier version of this manuscript. He would also like to thank the referees for helping him to clarify and reorganize the material.

REFERENCES

BLACKWELL, D. (1951). Comparison of experiments. *Proc. Second Berkeley Symp. Math. Statist. Probab.* 93–102. Univ. California Press.

BLACKWELL, D. and GIRSCHICK, M. A. (1954). *Theory of Games and Statistical Decisions*. Wiley, New York.

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

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

DAWID, A. P. (1985). Calibration-based empirical probability. *Ann. Statist.* **13** 1251–1274.

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

DEGROOT, M. H. and FIENBERG, S. E. (1982b). The comparison and evaluation of forecasters. Technical Report 244, Dept. Statistics, Carnegie Mellon Univ.

EPSTEIN, E. S. (1962). A Bayesian approach to decision making in applied meteorology. *J. Appl. Meteorology* **1** 169–177.

HENDRICKSON, A. D. and BUEHLER, R. J. (1971). Proper scores for probability forecasters. *Ann. Math. Statist.* **42** 1916–1921.

HEWITT, E. and STROMBERG, K. (1965). *Real and Abstract Analysis*. Springer, New York.

LINDLEY, D. V. (1982). The improvement of probability judgements. *J. Roy. Statist. Soc. Ser. A* **145** 117–126.

LINDLEY, D. V., TVERSKY, A. and BROWN, R. V. (1979). On the reconciliation of probability assessments (with discussion). *J. Roy. Statist. Soc. Ser. A* **142** 146–180.

MURPHY, A. H. (1972). Scalar and vector partitions of the probability score: Part I. Two-state

situation. *J. Appl. Meteorology* **11** 273–282.

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. and EPSTEIN, E. S. (1967). A note on probability forecasts and "hedging." *J. Appl. Meteorology* **6** 1002–1004.

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

OAKES, D. (1985). Self-calibrating priors do not exist. *J. Amer. Statist. Assoc.* **80** 339.

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

SCHERVISH, M. J. (1985a). Discussion of "Self-calibrating priors do not exist," by D. Oakes. *J. Amer. Statist. Assoc.* **80** 341–342.

SCHERVISH, M. J. (1985b). Discussion of "Calibration-based empirical probability" by A. P. Dawid. *Ann. Statist.* **13** 1274–1282.

SHUFORD, E. H., ALBERT, A. and MASSENGILL, H. E. (1966). Admissible probability measurement procedures. *Psychometrika* **31** 125–145.

THOMPSON, J. C. (1952). On the operational deficiencies in categorical weather forecasts. *Bull. Amer. Meteorological Soc.* **33** 223–226.

THOMPSON, J. C. (1962). Economic gains from scientific advances and operational improvements in meteorological prediction. *J. Appl. Meteorology* **1** 13–17.

THOMPSON, J. C. and BRIER, G. W. (1955). The economic utility of weather forecasts. *Monthly Weather Review* **83** 249–254.

VARDEMAN, S. and MEEDEN, G. (1983). Calibration, sufficiency, and domination considerations for Bayesian probability assessors. *J. Amer. Statist. Assoc.* **78** 808–816.

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

WINKLER, R. L. and MURPHY, A. H. (1979). The value of weather forecasts in the cost-loss ratio situation: An ex ante approach. In *Preprints Sixth Conf. Probab. Statist. Atmospheric Sciences* 134–138. Amer. Meteorological Soc., Boston.

DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213