

A Signal Detection Framework for the Evaluation of Probabilistic Forecasts

KEITH LEVI

University of Michigan

In this paper I formulate an approach for evaluating probabilistic forecasts in terms of signal detection theory. Signal detection theory provides a powerful perspective for this type of problem, and a rich empirical background including methodological tools as well as an extensive body of research in many domains. I propose procedures which emphasize the maximization of expected utility for the decision maker who uses the forecasts. Further, I suggest approaches to obtaining indices of calibration and resolution within this framework. I also present arguments that the proposed indices will exhibit the same basic properties as do decompositions of Brier's (1950, *Monthly Weather Review*, 78, 1-3) mean probability score. However, the properties may be reflected in different ways, and hence, the present methods may lead to different conclusions about forecasting ability. Finally, I argue that the use of an expected utility loss function makes this approach more appropriate for practical applications as well as for theoretical research than other procedures with more arbitrary loss functions. © 1985 Academic Press, Inc.

Probabilistic inferences and forecasts are often a critical component of any decision making situation. Decisions to buy or sell, operate or not operate, attack or negotiate, buy insurance or take a risk, bet on the Yankees or the Dodgers, and an endless number of other decisions, ranging from routine daily decisions of individuals to world-shaping decisions of societies and governments, will often depend on inferences and predictions about uncertain or future states of the world. In many situations the decision maker seeks the advice or opinion of an expert in regard to these uncertain states of the world. This paper is concerned with the evaluation of such forecasts from the perspective of the decision maker who wants to utilize the forecasts for making a decision.

The evaluation procedures to be presented here will emphasize the practical value of the forecasts for the decision maker who uses them. This emphasis leads to the use of expected utility loss functions. The particular emphasis of this paper will be to develop expected utility eval-

I thank J. Frank Yates and William M. Goldstein for the many discussions which led to both the formation and development of many of the ideas in this paper, and for their critical comments on numerous drafts. I also thank two anonymous reviewers for many valuable comments and suggestions. This research was supported in part by an NSF Graduate Fellowship to the author. Requests for reprints should be addressed to the author at Human Performance Center, 330 Packard Road, University of Michigan, Ann Arbor, Michigan 48104-1346.

uation procedures in the signal detection paradigm. Signal detection theory offers a powerful conceptual perspective and many empirical tools for this type of problem. I will further argue that the use of such loss functions is not only important for practical applications but also for theoretical concerns about underlying processes and abilities involved in probabilistic forecasts.

The evaluation procedures will focus on two main issues. The first is how should a given decision maker choose a forecaster from whom predictions will be obtained. The second is how can different skills involved in making probabilistic predictions be characterized. Two specific skills will be considered. One is the ability to place events into distinct categories depending on their outcome. For example, one wants a weather forecaster to be able to accurately separate weather into categories such as "Rain" or "No Rain." The second skill is forecasters' ability to accurately quantify this uncertainty. For example, one expects that there will be rain on a large majority of the days for which there was a 90% chance of rain forecast.

An issue with which this paper is not explicitly concerned is coercing the forecaster to give honest forecasts. This contrasts with the mainstream literature on scoring rules for probabilistic forecasts which deals principally with "proper" scoring rules. Proper scoring rules are those for which forecasters can maximize their expected score only by giving an honest forecast. It has long been recognized that this property is only relevant to the elicitation of forecasts, not their evaluation (Murphy & Winkler, 1970; Winkler, 1969). Therefore, the popularity of proper rules as evaluation measures is probably due to their decomposition properties rather than their properness. That is, some of these scoring rules can be decomposed into terms which reflect the forecasting skills mentioned above. For example, Sanders (1963), Murphy (1973), and Yates (1982) have all developed decompositions of a particular rule, the "mean probability score (\overline{PS})" (Brier, 1950). This rule is simply a squared error loss function of predictions and outcomes. Specifically, the score is the average of the difference between the forecast probability and a 0-1 outcome index. For example, suppose a weather forecaster predicted 30, 60, and 90% chances of rain on 3 consecutive days. If it rained on the first and third days the mean probability score would be $[(.3-1)^2 + (.6-0)^2 + (.9-1)^2]/3 = .287$. A smaller score is better.

Sanders (1963) and Murphy (1973) have shown that \overline{PS} can be partitioned into terms reflecting the concepts of resolution and calibration. According to these decompositions, a well-resolved forecaster never assigns the same probability to events with different outcomes. For example, suppose all the occasions on which an event occurred had been assigned probabilities of .1, .5, or .9, and all the occasions on which it

did not occur were assigned probabilities .2, .6, or .8. This forecaster would receive a perfect resolution score.

Whereas resolution is reflective of a forecaster's substantive expertise in the sense of distinguishing between outcomes, calibration is a measure of how adept the forecaster is at assigning numerical probabilities. In the traditional sense, a well-calibrated forecaster assigns probabilities such that, for a set of equally likely events, the proportion of events which occur is the same as the assigned probability.

Although these decompositions provide interesting indices of forecasting abilities, it is not at all clear that a decision maker should use a rule such as \overline{PS} for evaluating forecasts and selecting among forecasters. The problem is that the squared error loss function of \overline{PS} may simply not be optimal for different decision makers. Different individuals undoubtedly have different attitudes and preferences about the type and magnitude of errors they are willing to tolerate.

The procedures of this paper will take such individual differences into account, and they will also provide indices of resolution and calibration skills. These indices are closely related to the traditional indices of \overline{PS} decompositions. However, certain unreasonable situations, such as the earlier resolution example, are eliminated. Another difference is that the present approach combines these indices into the total evaluation so as to maximize the decision maker's utility. Thus, it is possible that the resolution and calibration indices a forecaster receives from the present procedures and \overline{PS} decompositions could be very similar, but the overall evaluations could be very different.

The importance of using a suitable loss function for empirical studies of forecasting was illustrated in a study of self-judged knowledge (Yates, 1982). This study found that people's self-judged knowledge about the events in concern was a very poor predictor of the external correspondence of their forecasts as measured by \overline{PS} . The question remains whether the subjects actually had poor self-insight, or whether \overline{PS} was simply a poor measure of external correspondence. It is hypothesized that the current evaluations would be much more predictive of self-judged knowledge. There was evidence in the Yates study that subjects with little self-judged knowledge were fortuitously well-calibrated (their forecasts of .5 turned out to be a good estimate of the base rate), whereas subjects with more self-judged knowledge were not well-calibrated in terms of the base rate. Since base-rate calibration is a factor in \overline{PS} , these effects may have obscured the ability of subjects with higher self-judged knowledge to differentiate between occurrences of the different outcome events.

Indeed, a reanalysis of the Yates data by the present author showed that the covariance term S_{fd} , of the decomposition was predictive of self-

judged knowledge. This term is an important factor in the present signal detection evaluations—it measures the mean separation of the distributions.

A study by Morris (1982) illustrates the significance of a suitable loss function in an applied context. He found that, although subjects scored poorly according to a calibration measure, they did quite well according to a certain operational evaluation. Thus, the above studies demonstrate that it is important to have a suitable loss function for both applied and theoretical evaluations.

There are several groups of papers which have dealt with expected utility evaluations for scoring rules. In the 1950s, meteorologists minimized expected monetary loss over a series of decisions about whether or not to prepare for adverse weather (Gringorten, 1959; Thompson, 1952; Thompson & Brier, 1955). Murphy (1966, 1969a, 1969b, 1976, 1977) has published a series of papers dealing with the relationships of proper scoring rules and expected utility. In particular, Murphy (1966) shows that a certain “expected utility” measure is equivalent to \overline{PS} . Such a result would justify the use of \overline{PS} for selecting which forecaster to employ. However, Murphy’s expected utility measure is a very restricted case because it is assumed to be linear with monetary losses and is averaged over a number of situations. Hence, it is really an expected value measure, similar to the earlier work. Further, Murphy’s derivation of equivalence involves assuming that the losses which are to be minimized are uniformly distributed random variables. In contrast, the present evaluation methods use general von Neumann and Morgenstern (N-M) utilities and are applicable to specific utility functions as well as random or average utility functions.

Savage (1971) proposed what he called the “share-of-the business” scoring rule. The basic idea of this rule was to give the forecaster a percentage of the profits (or losses) that the decision maker incurred. Closely related to this idea is a large body of literature in economics known as “agency theory” (e.g., see Grossman & Hart, 1983; Harris & Raviv, 1978; Stiglitz, 1974). This theory involves a principle-agent model where the principle (or decision maker in the present context) wants some output (e.g., a forecast) from the agent. Agency theory deals with the design of optimal incentive schemes which take into account the decision maker’s utility function and the agent’s costs for providing the desired service. Thus, both the share-of-the-business rule and agency theory involve expected utility considerations. However, neither is directly concerned with evaluating forecast quality, but rather they deal with forecast elicitation.

More closely related to the present paper are previous papers where forecasts are revised in accordance with Bayes’ rule (Lindley, 1982; Lind-

ley, Tversky, & Brown, 1979; Morris, 1974, 1977). Lindley's approach (1982), especially the frequency evaluation, is most closely related to this paper. Both involve frequency distributions of forecasts as functions of dichotomous outcomes. The main contribution of this paper is to put the problem in a signal detection perspective.

The paradigmatic problem of signal detection theory (Green & Swets, 1974) is the decision of whether an observation came from one of two distributions. Green and Swets (1974) showed that the basic likelihood ratio decision rule of signal detection theory has as special cases almost all commonly used decision rules. In particular, depending on the choice of a constant, the likelihood ratio rule will maximize any weighted combination of hits and false alarms, percentage correct, the Neyman-Pearson objective, and of course, expected utility as in the decision analysis paradigm. The value of signal detection theory in the present context is that it offers a powerful perspective and language for this type of problem. Further, a large body of empirical tools and research has been developed in this paradigm over the last 20 years. The goal of the present paper is to apply some of these tools for the evaluation of probabilistic forecasts, and in particular to propose procedures for obtaining indices of resolution and calibration within this framework.

In an independent development, an Australian meteorologist (Mason, 1982) has also come up with the idea of using a signal detection framework for evaluating probabilistic forecasts. Mason's concern was mainly to have an index of performance which was independent of the numerical properties of the forecasts—i.e., one that ignored the calibration of the forecasts. In contrast, the approach of this paper makes great use of the actual numbers used by the forecaster. Essentially, Mason proposes what is equivalent to the third procedure of this paper—using the area under the ROC curve as an index of resolution. He does not include the expected utility justifications of the present paper, but he does present empirical evidence concerning normality of underlying distributions. Those data argue for the practicality of implementing the proposed signal detection-based procedures.

SIGNAL DETECTION THEORY FRAMEWORK

A basic assumption of signal detection theory (Green & Swets, 1974) is that the performance of the detector can be explained in terms of two underlying distributions. In the classical situation one distribution represents random noise, and the other represents random noise with a signal added to it. A natural distribution for random noise is the normal distribution. Since the other distribution is just noise with a signal added, it is also assumed to be normal with the same variance but a different mean.

The present procedures also postulate two underlying distributions, but

they are not assumed to be normal. However, it will be assumed that the two distributions are on a scale which is monotonic in the likelihood ratio of the distributions. This is also a basic assumption in the standard signal detection paradigm. In the present case the underlying scale which is monotonic with the likelihood ratio will be called a "strength of evidence" scale. One distribution will represent the perceived strength of evidence which was present when an event occurred, and the other will represent the perceived strength of evidence when the event did not occur.

It will also be assumed that the forecaster reports a higher probability for the event with the larger likelihood ratio. Thus, since the likelihood ratio has been assumed to be monotonic with strength of evidence, the above assumption is equivalent to saying that the event with greater strength of evidence will always be given a higher forecast probability. For example, if one were predicting the outcomes of baseball games in terms of home teams winning, then the games for which the evidence most favored the home team would receive the higher probabilities.

In the classical signal detection paradigm the detector must report "Yes, there is a signal" or "No, there is no signal, only noise." There are four possible results: (a) a signal is reported and actually occurred (called a "hit"); (b) a signal is reported even though it had not been present (called a "false alarm" (FA)); (c) a signal is not reported, nor was one present (called a "correct rejection" (CR)); (d) a signal was not reported, although one was present (called a "miss").

Note that, of the times a signal was present, the proportion of hits is equal to one minus the proportion of misses. The same relationship holds for FAs and CRs when a signal was not present. Thus, we can summarize this information as a point in a two-dimensional space with proportion of hits on the vertical axis and proportion of FAs on the horizontal axis. Other points can be located in this space by altering the detector's response criterion. The curve which is created by connecting these points is known as the receiver operating characteristic (ROC) curve. Thus, the ROC curve shows the proportion of FAs which accompanies any level of hits.

It is well known that ROC curves can also be generated directly from rating responses—which are essentially what a forecaster reports. An ROC curve is generated from a rating or forecast response by computing the proportions of hits and FAs which would occur for all likelihood ratio decision criteria. For example, suppose action A_1 was taken only if the probabilistic forecast was .99 or above. This would generate one point on the ROC curve—with lots of misses and CRs, undoubtedly. Other points would be generated for the criteria of .98 and above, .97 and above, .96 and above, etc. Thus, the proportions of hits and FAs are computed

		ACTIONS	
		A ₁ "YES"	A ₂ "NO"
OUTCOMES	O ₁ (SIGNAL)	HIT	MISS
	O ₂ (NOISE)	FA	CR

FIG. 1. Outcome-action matrix showing the four possible outcomes of a dichotomous decision. Action A_1 corresponds to saying "yes," A_2 to saying "no." Outcome O_1 corresponds to a signal being present, O_2 corresponds to noise.

by assuming that any event assigned a higher probability than the threshold would result in action A_1 being taken, and any event below the threshold would result in action A_2 being taken. Therefore, in addition to the earlier assumptions, it will also be assumed that decision makers follow a likelihood ratio decision rule in the sense that if they take action A_1 for some event, then they will also take action A_1 for any event with a higher forecast probability. This will ensure the equivalence of the rating and yes-no ROC curves.

A basic property of signal detection theory is that it allows one to separate discrimination abilities from decision making biases. This property is directly utilized here in that two different people are used for the two tasks. The forecaster acts as the signal detector and produces the ROC curve. As will be seen later, the decision maker can then supply the utilities which determine the point of operation on the ROC curve.

Even without assuming normal distributions, there are some desirable expected properties of the ROC curves. First, recall the assumptions that the strength of evidence scale and the likelihood ratio of the two distributions are monotonically related, and that a likelihood ratio decision rule is used. Using these assumptions it is easy to demonstrate that the slope of the ROC curve is equal to the likelihood ratio. This can be shown (e.g., Green & Swets, 1974, p. 38) by taking derivatives of the coordinates of the ROC curve. These derivatives are equal to the respective likelihoods. Thus, the ratio of the derivatives gives the slope and the likelihood ratio. Under these conditions, it is also well known that an ROC curve must have a hit probability which is a monotonically increasing function of the FA probability, and a slope that is monotonically decreasing.

In summary, I have postulated that forecasters' behavior can be described as if they mapped their information about events onto a strength

of evidence scale. For a given forecaster, two distributions are created on this scale depending on which of the two possible outcomes occurs for each event. The basic assumptions about this scale and the distributions are:

(i) The scale is monotonic with the likelihood ratio of the two distributions.

(ii) A forecaster's reported probabilities are monotone with the likelihood ratios.

(iii) A likelihood ratio decision rule is used.

These assumptions have implied several necessary properties of the ROC curves which will be useful in the following evaluation procedures.

EVALUATION PROCEDURES

Three different evaluation procedures which use the signal detection framework will be presented next. Each is based on the maximization of expected utility (EU) for the user of the forecasts. The first and second are designed to apply to individual decision makers. The third can be used when considering a large group of decision makers with diverse or unknown utility functions. The first and third principally reflect the concept of resolution. The second involves the concept of calibration in addition to resolution.

I will first provide summary descriptions of the three evaluation procedures. Then, each procedure will be discussed in some detail. All three procedures assume that probabilistic forecasts have been given for a set of dichotomous events, and that the outcomes have been observed.

Optimal EU Evaluation

(a) Trace out the ROC curve for each forecaster, and find the optimal operating point on the curve.

(b) Estimate the probability of a hit, miss, FA, or CR at this point.

(c) Calculate the expected utility.

Face-Value EU Evaluation

(a) For each forecaster, estimate the probability of a hit, miss, FA, or CR using the decision rule:

Act on O_1 if

$$r \geq (U_{22} - U_{12}) / [(U_{22} - U_{12}) + (U_{11} - U_{21})] \quad (1)$$

where r is the forecasted probability of O_1 occurring, and U_{ij} is the decision maker's utility for outcome i when he acts for outcome j .

(b) Calculate the expected utility.

Area under the ROC Curve

Trace out the ROC curve and find the area beneath it.

Expected Utility Evaluations

Basic principles and common assumptions. The basic principle underlying the first two evaluation methods is the maximization of EU for the decision maker, in the following sense. Suppose a decision maker wants to choose one of several forecasters to predict the outcome of some event. If the forecasters have predicted outcomes for a number of similar events in the past, then ROC curves can be defined for each forecaster and the probabilities of a hit, miss, FA, or CR can be estimated for any particular decision rule. The decision maker's EU for using each forecaster can then be calculated from these probabilities. These EUs are the forecasters' scores

Distinction between optimal and face-value EU scores. Now, note that the estimated probabilities of hits and FAs depend on where one is on a forecaster's ROC curve. That is, they depend on what decision rule is used. The two different EU evaluation procedures will correspond to the use of different decision rules by the decision maker. One rule will be to find the point on the obtained ROC curve which maximizes the decision maker's EU. This rule will be called the "optimal EU evaluation rule." The other will be to use the forecaster's reported probabilities at face value, and will be called the "face-value EU evaluation rule."

The decision maker takes forecasts at face value by using the forecaster's reported probabilities in computing whether the EU of taking act A_1 or A_2 is greater. This is in contrast to finding the optimal point on the ROC curve which implies the use of a decision rule which may not be the same as the face-value rule. The possible nonequivalence of these decision rules is demonstrated next.

As I pointed out earlier, Green and Swets (1974) have shown that in the case of two-event outcomes, a likelihood ratio decision rule of the following form will maximize a number of decision criteria: "Prepare for outcome O_1 if and only if the likelihood ratio is greater than some constant." It is instructive to review the derivation of the decision threshold which maximizes EU and note its interpretation in the present situation. First, let

$$\begin{aligned} U_{ij} &= \text{utility of taking action } i \text{ when outcome } j \text{ occurs,} \\ r &= \text{reported probability that outcome } O_1 \text{ will occur.} \end{aligned}$$

Then, given the reported forecast, the decision maker should take action A_1 if and only if

$$\begin{aligned}
& EU(A_1|r) \geq EU(A_2|r) \\
\Leftrightarrow & p(O_1|r)U_{11} + p(O_2|r)U_{12} \geq p(O_1|r)U_{21} + p(O_2|r)U_{22} \\
\Leftrightarrow & p(O_1|r)[U_{11} - U_{21}] \geq p(O_2|r)[U_{22} - U_{12}] \\
\Leftrightarrow & p(O_1|r)/p(O_2|r) \geq \beta \tag{2} \\
\Leftrightarrow & p(r|O_1)/p(r|O_2) \geq \beta[p(O_2)/p(O_1)] \tag{3}
\end{aligned}$$

where

$$\beta = [U_{22} - U_{12}]/[U_{11} - U_{21}].$$

If the decision maker takes the forecaster's probabilities at face value then $p(O_j|r)$ is equal to the forecaster's reported posterior probability of outcome j occurring, given the forecaster's perceived strength of evidence. That is, $p(O_1|r) = r$ and $p(O_2|r) = 1 - r$. Thus, (2) gives a decision rule based on posterior probabilities. Using Bayes' Theorem, this can be transformed to a rule using the likelihood ratio and prior probabilities, as in (3). Thus, (2) and (3) specify decision rules of the form, "Take A_1 if and only if the ratio of posterior probabilities is at least as large as the decision threshold on the right side of (2), or if and only if the likelihood ratio is at least as large as the decision threshold on the right side of (3)."

Now recall that the earlier assumptions about the strength of evidence distributions imply that the slope of the ROC curve is equal to the likelihood ratio. Therefore, the point on the ROC curve where the slope equals the decision threshold in (3) gives the decision rule for which the decision maker will maximize expected utility.

The question here is whether the rule implied by the optimal point on the ROC curve is equivalent, in the sense of producing the same outcomes, as the rule the decision maker would follow if he took the forecasts at face value. The above derivation indicates a formal equivalency. However, due to hedging or miscalibration, the forecaster's reported posterior probabilities may not be equivalent to those implied by the ROC curves and the underlying distributions. A numerical example of such a case is presented later in this paper. In this situation, if forecasts are taken at face value, the decision maker operates at a different point on the ROC curve than is optimal. Thus, in order to do as well as possible in the face-value EU evaluation, a forecaster needs to be honest and well-calibrated in the sense that predictions accurately reflect underlying strength of evidence distributions.

Note that this argument might constitute an expected utility justification for using a proper scoring rule to evaluate forecasts. That is, if forecasts accurately reflect posterior probabilities, then the decision maker would maximize his expected utility for that forecaster. Of course, it is not necessarily true that forecasts will be well calibrated in the sense of

reflecting the probabilities implied by the ROC curves even if they are honestly given.

Although the face-value EU evaluation depends on calibration, the optimal EU rule is mainly a measure of the concept of resolution. It ignores the numerical calibration properties of the forecasts in the sense that it only makes use of the order relations. Hence, it reflects the ability to order the events so that the events for which outcome O_1 occurs are all ranked above those events for which outcome O_2 occurs.

Another way to think of the two EU evaluation procedures is in terms of an analogy with the theory of the ideal receiver (Green & Swets, 1974). Both evaluation rules are based on the same ROC curve and the same underlying distributions. The optimal EU evaluation acts like an ideal receiver in that it finds the optimal operating point on the ROC curve. The face-value EU evaluation finds a point on the ROC curve which may or may not be the optimal point. Thus, comparisons of the optimal EU evaluation with the face-value EU evaluation yield indications of the forecaster's accuracy in assigning numbers to represent their knowledge of the situation.

Computing EU for a future event. The actual score yielded by each of these rules is simply the EU for using a forecaster's prediction of a future event.

Let

U_{ij} = decision maker's utility when action A_i is taken and outcome O_j occurs, $i, j = 1, 2$, and

P_{ij} = the joint probability that action A_i will be taken and outcome O_j will occur, $i, j = 1, 2$.

Thus, the decision maker's expected utility for the subsequent decision is

$$EU = \sum_{i=1}^2 \sum_{j=1}^2 P_{ij} U_{ij} \quad (4)$$

Now let

P_{ij} = probability that action A_i is taken conditional on outcome O_j occurring, and

P_j = probability of outcome O_j occurring.

Hence, P_{ij} is simply the probability of a hit, FA, miss, or CR, depending on i and j . Obviously, these probabilities vary over the ROC curve and therefore depend on what decision rule is used. The probabilities are estimated by the appropriate proportions of obtained outcomes for particular decision rules. The marginal probabilities P_1 and P_2 can be esti-

mated by the observed proportions of O_1 and O_2 . Thus, P_{ij} , P_j , U_{ij} are all estimable quantities, and the expected utility is easy to compute. In particular,

$$EU = \sum_{i=1}^2 \sum_{j=1}^2 P_{ij} P_j U_{ij}$$

and this is estimated by

$$\begin{aligned} EU^* &= \sum_{i=1}^2 \sum_{j=1}^2 (n_{ij}/n_j)(n_j/N)U_{ij} \\ &= \sum_{i=1}^2 \sum_{j=1}^2 (n_{ij}/N)U_{ij}, \end{aligned} \quad (5)$$

where

N = total number of events,

n_j = number of events for which O_j obtained, and

n_{ij} = number of events for which O_j obtained and action A_i was taken.

Dominance in the optimal EU evaluation. In many situations it might not even be necessary to know a decision maker's utilities in order to determine which forecaster will maximize the optimal EU score. It is possible that the same ordering of the forecasters would hold for all decision makers. Suppose the ROC curve of one forecaster dominates the curve of another, as in Figure 2. Note that, for every point on curve B, there is a point from curve D directly above it and a point directly to the left of it. That is, for every point on curve B, there is a point on curve D with more hits and fewer FAs. It can be shown that EU will always be maximized for the decision maker who uses judgments of the forecaster with the dominant ROC curve and the rule determined by the point on the curve where the slope equals the decision threshold. Intuitively, suppose that points b and d are where the slope equals the decision threshold on curves B and D, respectively. Point d does not directly dominate b . However, f , g , and all points between them on D do dominate b . Since d maximizes the expected utility of all points on D, it must also do better than f and g , and therefore b . This argument uses conditional probabilities. To make the argument rigorous, it must be shown that it also holds for unconditional probabilities, as in (4). A formal proof is given in the Appendix.

Thus, when the ROC curve of one forecaster dominates that of another forecaster, one does not even need to know the decision maker's utilities in order to determine the optimal EU evaluation. The forecaster with the

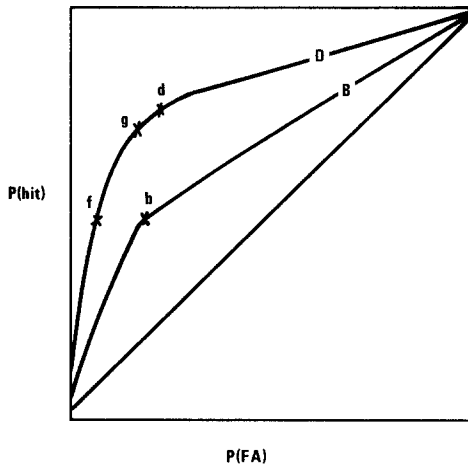


FIG. 2. Illustration of a dominant ROC curve. If point d dominates all points on curve D in terms of expected utility, then it must also dominate all points on curve B .

dominant ROC curve will be better for any decision threshold. Thus, the optimal EU evaluation is very simple and clearcut when dominance holds.

The dominance condition only guarantees the evaluation ordering for the optimal EU evaluation, not the face-value EU evaluation. This is because if the forecaster reports probabilities which are not equivalent to those implied by the ROC curve, then the decision maker will not be operating at the optimal point on a curve. Hence, one could operate on a dominated ROC curve and still obtain the highest expected utility.

Of course, it is an empirical question whether dominance will be found. However, it seems that in most situations it will not be unreasonable to expect to find dominance. For example, in situations where the event outcomes are very similar—for example, win or lose a sporting event—it is reasonable to expect both distributions to be approximately normal with similar variances. These curves should tend to be well behaved.

In particular, if distinct ROC curves have normally distributed equal variance distributions, then they cannot intersect each other. If they did intersect, then there would be a point where the probabilities of a hit and FA for one curve equaled those of the other. If the distributions have equal variances, then such a point would constrain the two sets of distributions to have the same distance between their means. Hence, the ROC curves could not be distinct.

Of course, this result assumes exactly equal variance normal distributions. Although these conditions are never perfectly satisfied in empirical situations, the equal variance normal ROC model has proven to be very robust in a wide variety of applications (e.g., see Swets & Pickett,

1982). Indeed, Mason (1982) found the model to provide a close fit to data from a large number of weather forecasting studies. Of particular interest is the relatively good fit of the equal variance normal model to the data from Murphy and Winkler's (1977) forecaster "B." This result of Mason is striking because a covariance decomposition analysis (Yates, 1984) of this forecaster appears to imply substantial violations of the normal equal variance assumptions.

Another common situation may arise where one outcome is much less variable than the other. For example, it might be that conditions for not rain are very unequivocal and therefore on the days that it does not rain might be quite accurate. On the other hand, the conditions which lead to rain might be much more ambiguous, and therefore these forecasts could have a high variance (e.g., Yates, 1984). In such situations all ROC curves should be skewed the same way and again dominance might be a common occurrence.

Situations which could possibly cause trouble would be when different ROC curves are skewed in different directions. Such a situation might arise if one forecaster was good at predicting one outcome and another forecaster was good at predicting the complementary outcome. It is not easy to imagine such situations, but if they should occur it makes sense that the choice of the better forecaster should depend on the decision maker's tradeoffs between different types of errors.

Area under the ROC Curve

The third evaluation procedure is to simply find the area under the ROC curve. This is essentially the approach pursued by Mason (1982). However, he argues that, rather than using the area under the nonparametric ROC curve, one should assume normal distributions underlying the ROC and use a parametric measure. In particular, he recommends A_z , which is the area under the ROC curve when the ROC curve is determined by a best fit assuming normality. The important point is that this is a generalized measure of resolution in the sense that it takes all utility functions into account. It measures resolution in the sense that the area under the ROC curve is mathematically a measure of the separation of the entire masses of the two underlying distributions. It considers all utility functions in the sense that each point on the ROC curve corresponds to a different decision rule, or a different utility function.

When the dominance condition holds, this evaluation rule will be ordinarily equivalent to the optimal EU rule. In such situations, the above results on the optimal EU evaluation justify the use of the area under the ROC in terms of EU theory. That is, when dominance holds, every decision maker will maximize EU by using the judgments of the forecaster

TABLE 1
NUMBER OF OBTAINED OUTCOMES FOR EACH FORECAST CATEGORY

Forecast ^a	Physician A		Physician B	
	Mal.	Ben.	Mal.	Ben.
.1	20	40	2	10
.3	40	70	5	20
.5	40	40	20	50
.7	20	5	48	40
.9	5	0	50	35

^a Forecast gives the physicians' reported subjective probability that a tumor was malignant. Mal. and Ben. give the number of cases which turned out to be malignant or benign.

with the dominant ROC curve, or equivalently, the judgments of the forecaster with the larger area under the ROC curve.

When the dominance condition does not hold, there are two views to consider. The first, which is the perspective of this paper, is that one should be very cautious in using the area under the ROC curve without dominance because different utility functions will be maximized on different curves. In this case the face-value or optimal EU evaluations should be used.

The alternative view is to focus on the area under the ROC curve as an accuracy index and use the standard justification from signal detection theory. In particular, Green and Swets (1974) showed that the area under the ROC curve is equivalent to the expected percentage correct in a two-alternative forced-choice experiment. For example, suppose one were interested in the outcomes of baseball games. A forced-choice task would be to present forecasters with two games and ask them to choose the game in which they felt the home team was most likely to win. Green and Swets (1974) argue that since forced-choice tasks always require a decision, they are essentially pure detection tasks. Hence, they are an intuitively justifiable measure of detection accuracy. Thus, when dominance fails, the area under the ROC curve is justifiable as an index of detection accuracy, but it does not guarantee the maximization of expected utility.

An Example

Imagine that you recently discovered a lump in your chest and would like to have a specialist read the X rays to diagnose whether the tumor is malignant or benign. Suppose Physicians A and B are the available specialists and you would like to evaluate the accuracy of their diagnoses. Suppose that data are available on past diagnoses of the two physicians as in Table 1. Table 1 gives the subjective probabilities of malignancy reported by the physicians, and the observed outcomes for each category.

TABLE 2
HIT, MISS, FA, AND CR RATES CORRESPONDING TO OUTCOMES IN TABLE 1^a

Decision thrsld.	Hits		Misses		P(hit)		FA		CR		P(FA)	
	A	B	A	B	A	B	A	B	A	B	A	B
.1	125	125	0	0	1.0	1.0	155	155	0	0	1.0	1.0
.3	105	123	20	2	.84	.98	115	145	40	10	.74	.94
.5	65	118	60	7	.52	.94	45	125	110	30	.29	.81
.7	25	98	100	27	.2	.78	5	75	150	80	.03	.48
.9	5	50	120	75	.04	.40	0	35	155	120	0.0	.23
1.0	0	0	125	125	0.0	0.0	155	155	0	0	0.0	0.0

^a Each row gives the rates of hits, misses, FAs, and CRs for physicians A and B corresponding to the decision rule, "Take treatment for a malignant tumor if the physician reports a likelihood of malignancy equal to or greater than the threshold."

The possible frequencies of hits, misses, FAs, and CRs are listed in Table 2. These results produce the ROC curves shown in Figure 3 for Physicians A and B.

Now suppose that you consider a FA to be equally aversive as a miss. For example, you think it would be equally bad to be operated on unnecessarily as it would be to ignore a malignant tumor. Further, suppose you consider treating a malignant tumor equally beneficial to not treating a benign tumor. In such a case, we might not find utilities: $U_{11} = U_{22} = 1$ and $U_{12} = U_{21} = 0$. Then, according to the optimal EU evaluation procedure, you should use the decision rule given by the decision threshold in (3). That is, be treated for cancer if the likelihood ratio is at least 1.24, since $\beta = 1$, $P_1 = .446$, and $P_2 = .554$. Since the slope of the ROC curve equals the likelihood ratio, you should operate at the points marked by the arrows in Figure 3. Thus, you should be treated for cancer

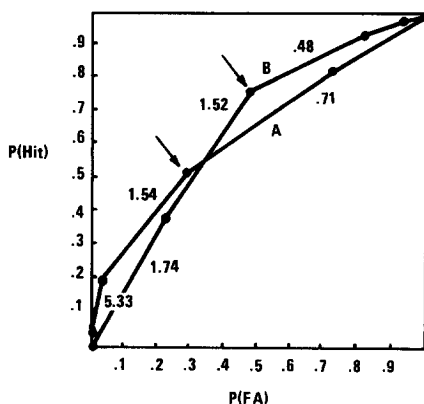


FIG. 3. ROC curves for physicians A and B of Tables 1 and 2. The numbers on the curves give the slopes of the adjacent line segments.

if Physician A says there is even a 50% chance the tumor is malignant. On the other hand, you should not be treated unless Physician B is at least 70% certain of malignancy. Using (5), the optimal EU for using Physician B is $[98(1) + 27(0) + 75(0) + 80(1)]/280 = .698$; for using Physician A it is $[65(1) + 60(0) + 45(0) + 110(1)]/280 = .625$. Thus, Physician B is favored by the optimal EU evaluation.

Now consider the face-value decision rule. According to the decision rule given by (1), you should be treated for a malignancy whenever the physician is more than 50% confident that the tumor is malignant. This is the same as the optimal EU rule for Physician A. Hence, the face-value EU for Physician A is again .625. However, this rule corresponds to a different point than does the optimal EU rule on Physicians B's ROC curve. Using this rule with Physician B, one would obtain 94% hits (or true positives) and 81% FAs (or false positives). Hence, the face-value EU for Physician B is $[118(1) + 7(0) + 125(0) + 30(1)]/280 = .529$.

Thus, Physician B produces a higher EU under the optimal EU procedure, but Physician A does better under the face-value EU procedure. Hence Physician A is better calibrated than Physician B in the sense that there is less of a loss in EU from the optimal to the face-value evaluation—at least in the current range of ROC curves. However, Physician B appears to be better resolved in this range. Put another way, for this particular patient, Physician B is potentially better at differentiating malignant tumors from benign tumors, but Physician A is better at quantifying the accuracy of diagnoses.

Now consider a patient for whom surgery or chemotherapy is very risky. A FA would be very bad for such a patient. Thus, utilities might be $U_{22} = 1$, $U_{11} = .7$, $U_{21} = .2$, and $U_{12} = 0$. The critical likelihood ratio is now 2.48. The optimal EU for Physician A is $[25(.7) + 100(.2) + 150(1) + 5(0)]/280 = .670$, and for Physician B it is $[0(.7) + 125(.2) + 155(1) + 0(0)]/280 = .643$. According to the face value EU rule, this patient should be treated if $r \geq .667$. Thus, the face value EU for physician A will again be .670, but for physician B it will be $[98(.7) + 27(.2) + 80(1) + 76(0)]/280 = .550$. Thus, this patient maximizes EU for both the optimal and face value EU rules by using physician A. These examples show that the best forecaster depends on the relative values of the particular decision maker. Such a result is not possible with traditional rules such as \overline{PS} .

To find the areas under the ROC curves one could simply connect the plotted points by straight lines, as in Figure 3, and find the areas under the lines. Or, if one assumes normal distributions, as in Mason (1982), there are computer programs which find the best fitting ROC curves and give measures of the areas under the curves (e.g., see Swets & Pickett, 1982).

DISCUSSION

I have presented three new evaluation procedures in the paradigm of signal detection theory. Signal detection theory provides a powerful perspective and a substantial body of empirical tools for this type of problem. In particular, this paper has emphasized the use of an expected utility loss function and the ability to obtain natural indices of calibration and resolution. The calibration and resolution indices presented here should be highly correlated with the traditional measures.

First, consider resolution. In this paper, resolution is measured by either the optimal EU evaluation or the area under the ROC curve. In either case, a high resolution score is obtained when there is some cutoff such that events which occur are assigned forecasts greater than the cutoff and events which do not occur are assigned forecasts below the cutoff. Thus, if we ignore the pathological cases such as assigning even forecasts to events which occur and odd forecasts to the others, then a forecaster is well resolved in the traditional sense precisely when that forecaster is well resolved in the present evaluations.

Now consider calibration. Let us call the traditional calibration measures "external calibration" indices. Thus, a forecaster is externally calibrated if for a set of events assigned a common forecast probability, the proportion which occur is equal to the forecast. I will call the calibration of this paper "internal calibration." Forecasters are internally calibrated to the degree they give forecasts resulting in decision makers operating at optimal points on the ROC curve.

Suppose a forecaster is not perfectly externally calibrated for some set of events assigned a common forecast probability. Then, holding other events constant, a decision maker whose decision cutoff is at this same point would end up operating at a point with more false alarms and hits than would be optimal (or fewer than optimal, depending on whether the forecast probability was greater or less than the outcome proportion). Hence, if one does not have perfect calibration, then neither does one have perfect internal calibration. That is, perfect internal calibration implies perfect external calibration.

On the other hand, if there is perfect external calibration, then the decision maker will be able to correctly choose the proper cutoff point, and hence will operate at the optimal point on the ROC curve, and hence will be perfectly internally calibrated. Thus, it appears that the measures of both resolution and calibration in the present evaluation procedures are directly related to the traditional measures.

Note that these evaluation rules, especially the optimal EU evaluation procedure and the area under the ROC, can be thought of as measuring two of the most important components of Yates' (1982) covariance de-

composition of \overline{PS} : (1) the separation of the two distributions of forecasts in terms of the difference between their means, and (2) the sum of their variances. However, whereas the covariance decomposition simply sums the variance and the covariance components, the signal detection scores evaluate the tradeoffs between these components in terms of finding the optimal combination of hits, misses, FAs, and CRs for the decision maker. Thus, although these evaluations reflect the same basic properties as do earlier decompositions of Brier's (1950) probability score, they may reflect them in different ways and thereby offer different insights.

The preceding arguments only discuss the signal detection calibration and resolution indices in general terms. Future work must investigate specific mathematical formulations of these indices. An obvious resolution index is the area under the ROC curve which, as previously noted, is a measure of the separation of the outcome distributions. A candidate for a general calibration measure is the sum of the difference of the face-value and optimal EU evaluations over all decision rules.

Such indices might alleviate a problem in decompositions of \overline{PS} . In particular, a reviewer of this paper pointed out that decompositions of \overline{PS} are surprisingly affected by whether one categorizes forecasts into fifths, tenths, hundredths, etc. These problems might be avoided in the signal detection framework. For example, fitting a curve as opposed to drawing straight lines between points along the ROC curve has been found to reduce problems in comparing ROC curves with different numbers of empirical points (Swets and Pickett, 1982). Thus, the investigation of specific signal detection indices should be a fertile topic for future studies.

Finally, let me note that the determination of a "suitable" loss function is not a closed question. The use of the N-M utility functions has been emphasized here because they allow the relative values of the decision maker to be taken into account, and they have been formally justified in terms of compelling axiom systems (e.g., see Coombs, Dawes, & Tversky, 1970; Luce & Raiffa, 1957; von Neumann & Morgenstern, 1953). Further, EU theory is very general and many other commonly used decision criteria can be instituted as special cases. Similar formal justifications have not yet been presented for the use of \overline{PS} and other commonly used scoring rules—other than the fact that they may be mathematically tractable and give some type of index of performance. When different indices of the same phenomena give different answers, then one must carefully consider the justifications underlying the different measures. The procedures of this paper emphasize the practical value of the forecasts to the person who uses them.

In conclusion, signal detection theory is a very elegant mathematical model and empirical framework with a rich history of theoretical and

empirical developments in many fields over the last 20 years. My goal in this paper has been to formulate the problem of evaluating probabilistic forecasts in terms of the signal detection paradigm and shown how indices of critical forecasting abilities can be developed in this context. The developments of this paper are by no means complete. Specific numerical indices and statistics must be developed for the procedures presented here. The precise mathematical relations between these indices and other scoring rules need to be determined. I am currently working on these problems. I am also planning studies for the near future which will utilize the evaluation procedures of this paper with real data and compare the results with traditional evaluations.

APPENDIX

The following proof shows that the unconditional utility defined in (4) is always greater for the point on a dominant ROC curve where the slope equals the likelihood ratio than it is for the unconditional utility at any point on a dominated ROC curve. I first show that the unconditional expected utility is always greater at a point which dominates another point in terms of probabilities of hits and FAs. After this, I show that the optimal likelihood ratio decision threshold does indeed give a higher expected utility than any other point on the same ROC curve, and therefore must also be greater than the expected utility at any point on a dominated ROC curve.

Without loss of generality, consider Figure 2 again. Since the dominating points on D all have more hits and fewer FAs than point b , $P_{11}^D > P_{11}^B$ and $P_{21}^D > P_{21}^B$, where P_{ij}^D refers to the relevant point on D and P_{ij}^B refers to point b . Note that $P_j^D = P_j^B$, $j = 1, 2$ since P_j is estimated by the proportions of observed outcomes O_j , which are the same for all forecasters—assuming that the evaluation is over the same set of events. Therefore, $P_{11}^D > P_{11}^B$ and $P_{22}^D > P_{22}^B$. Now note that

$$P_{21} = P_{21}P_1 = (1 - P_{11})P_1 = P_1 - P_{11}$$

$$P_{12} = P_{12}P_2 = (1 - P_{21})P_2 = P_2 - P_{22}$$

I also assume that $U_{11} \geq U_{21}$ and $U_{22} \geq U_{12}$. That is, the utility of the correct action is at least as good as the utility of the incorrect action for a given outcome. I can now show that the expected utility is at least as great for a dominating point as it is for a dominated point.

$$U_{11} \geq U_{21} \tag{6}$$

$$\Rightarrow U_{11}[P_{11}^D - P_{11}^B] \geq U_{21}[P_{11}^D - P_{11}^B]$$

$$\Rightarrow U_{11}[P_{11}^D - P_{11}^B] \geq U_{21}(P_1 - P_{11}^B) - (P_1 - P_{11}^D)$$

$$\Rightarrow P_{11}^D U_{11} - P_{11}^B U_{11} \geq P_{21}^B U_{21} - P_{21}^D U_{21}$$

$$\Rightarrow P_{11}^D U_{11} + P_{21}^D U_{21} \geq P_{11}^B U_{11} + P_{21}^B U_{21} \tag{7}$$

Similarly,

$$U_{22} \geq U_{12} \tag{8}$$

$$\Rightarrow U_{22}[P_{22}^D - P_{22}^B] \geq U_{12}[P_{22}^D - P_{22}^B]$$

$$\Rightarrow U_{11}[P_{11}^D - P_{11}^B] \geq U_{12}(P_2 - P_{22}^B) - (P_2 - P_{22}^D)$$

$$\Rightarrow P_{22}^D U_{22} - P_{22}^B U_{22} \geq P_{12}^B U_{12} - P_{12}^D U_{12}$$

$$\Rightarrow P_{22}^D U_{22} + P_{12}^D U_{12} \geq P_{22}^B U_{22} + P_{12}^B U_{12} \tag{9}$$

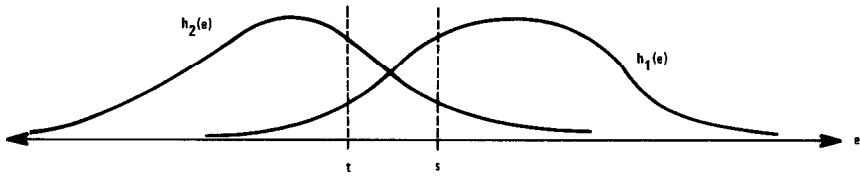


FIG. 4. Strength of evidence distributions. Points t and s correspond to points d and f , respectively, in Fig. 2.

Combining (7) and (9),

$$\sum_{i=1}^2 \sum_{j=1}^2 P_{ij}^D U_{ij} \geq \sum_{i=1}^2 \sum_{j=1}^2 P_{ij}^B U_{ij}.$$

And, the inequality is strict if either of inequalities (6) or (8) are strict. If neither (6) nor (8) is strict, then the decision problem is trivial because the expected utility will be the same for any forecaster. That is, $U_{11} = U_{21}$ and $U_{22} = U_{12}$, then let $U_1 = U_{11} = U_{21}$ and $U_2 = U_{22} = U_{12}$. Then, the result is the following identity:

$$\begin{aligned} \Rightarrow \quad & (P_{11}^D + P_{21}^D)U_1 + (P_{22}^D + P_{12}^D)U_2 = (P_{11}^B + P_{21}^B)U_1 + (P_{22}^B + P_{12}^B)U_2 \\ \Rightarrow \quad & \sum_{i=1}^2 \sum_{j=1}^2 P_{ij}^D U_{ij} = \sum_{i=1}^2 \sum_{j=1}^2 P_{ij}^B U_{ij}. \end{aligned}$$

Thus, it is safe to assume that if the hit and FA probabilities of one forecaster dominate those of another, then the expected utility to the decision maker is strictly greater at the dominating point.

Now I will show that the expected utility to the decision maker is greater at point d than at any other point on the curve D . That is, the point which maximizes expected utility conditioned on a given forecast also maximizes the unconditional expected utility. Without loss of generality, consider the point f on D . Both f and d correspond to unique points on the perceived strength of evidence scale. Call them s and t as in Figure 4. Thus, the slope at f in Figure 2 is equal to the likelihood ratio at s in Figure 4, and similarly for d and t .

If p_{ij}^d and p_{ij}^f are the conditional probabilities of taking action i given outcome j at points d and f of Figure 2, and $h_j(e)$, $j = 1, 2$, are the density functions for the perceived strength of evidence distributions in Figure 4, then,

$$\begin{aligned} P_{11}^d &= \int_t^x h_1(e)de = \int_t^s h_1(e)de + \int_s^x h_1(e)de = P_{11}^f + \int_t^s h_1(e)de \\ P_{21}^d &= \int_{-\infty}^t h_1(e)de = \int_{-\infty}^s h_1(e)de - \int_t^s h_1(e)de = P_{21}^f - \int_t^s h_1(e)de. \end{aligned}$$

Thus,

$$P_{11}^d = P_{11}^f + P_1 \int_t^s h_1(e)de \tag{10}$$

$$P_{21}^d = P_{21}^f - P_1 \int_t^s h_1(e)de. \tag{11}$$

Similarly,

$$P_{22}^d = \int_{-\infty}^t h_2(e)de = \int_{-\infty}^s h_2(e)de - \int_t^s h_2(e)de = P_{22}^f - \int_t^s h_2(e)de$$

$$P_{12}^d = \int_t^{\infty} h_2(e)de = \int_s^{\infty} h_2(e)de + \int_t^s h_2(e)de = P_{12}^f + \int_t^s h_2(e)de$$

$$P_{22}^d = P_{22}^f - P_2 \int_t^s h_2(e)de \tag{12}$$

$$P_{12}^d = P_{12}^f + P_2 \int_t^s h_2(e)de. \tag{13}$$

From the earlier derivation of the decision threshold in (3), it is known that t is the point at which

$$h_1(e)/h_2(e) = \beta(P_2/P_1),$$

where $\beta = [U_{22} - U_{12}]/[U_{11} - U_{21}]$ as before. Since the likelihood ratio is monotone increasing with e , it follows that

$$\Rightarrow \begin{aligned} & h_1(e)/h_2(e) > \beta(P_2/P_1) \quad \text{for all } e > t \\ & h_1(e) > \beta(P_2/P_1)h_2(e) \quad \text{for all } e > t \end{aligned}$$

$$\Rightarrow \int_t^s h_1(e)de > \beta(P_2/P_1) \int_t^s h_2(e)de$$

$$\Rightarrow \left[\int_t^s h_1(e)de \right] / \left[\int_t^s h_2(e)de \right] > \beta(P_2/P_1).$$

Rearranging terms,

$$\Rightarrow P_1 \int_t^s h_1(e)de[U_{11} - U_{21}] - P_2 \int_t^s h_2(e)de[U_{22} - U_{12}] > 0$$

$$\Rightarrow P_1 \int_t^s h_1(e)deU_{11} - P_1 \int_t^s h_1(e)deU_{21} - P_2 \int_t^s h_2(e)deU_{22} + P_2 \int_t^s h_2(e)deU_{12} > 0.$$

Substituting from (10), (11), (12), and (13), we obtain

$$(P_{11}^d - P_{11}^f)U_{11} + (P_{21}^d - P_{21}^f)U_{21} + (P_{22}^d - P_{22}^f)U_{22} + (P_{12}^d - P_{12}^f)U_{12} > 0$$

$$\Rightarrow \sum_{i=1}^2 \sum_{j=1}^2 P_{ij}^d U_{ij} > \sum_{i=1}^2 \sum_{j=1}^2 P_{ij}^f U_{ij}.$$

Thus, the expected utility is greater at point d than at point f . Note that f could be any point to the left of d on curve D . A symmetric argument will show that the decision threshold point, d , also has higher expected utility than any point to the right of it on curve D . I have now shown that a decision maker will always maximize expected utility by using the decision rule implied by the point on the dominant ROC curve where the slope equals the decision threshold.

REFERENCES

- Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**, 1–3.
- Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). *Mathematical psychology*. Englewood Cliffs, NJ: Prentice-Hall.
- Green, D. M., & Swets, J. A. (1974). *Signal detection theory and psychophysics* (2nd ed.). New York: Wiley.
- Gringorten, I. I. (1958). On the comparison of one or more sets of probability forecasts. *Journal of Meteorology*, **15**, 283–287.
- Grossman, S. J., Hart, O. D. (1983). An analysis of the principal-agent problem. *Econometrica*, **51**(1), 7–45.
- Harris, M., & Raviv, A. (1978). Some results on incentive contracts with applications to education and employment, health insurance, and law enforcement. *The American Economic Review*, **68**, 20–30.
- Lindley, D. V. (1982). The improvement of probability judgements. *Journal of the Royal Statistical Society*, **145**, 117–126.
- Lindley, D. V., Tversky, A., & Brown, R. V. (1979). On the reconciliation of probability assessments. *Journal of the Royal Statistical Society*, **142**, 146–180.
- Luce, R. D., & Raiffa, H. (1957). *Games and decisions*. New York: Wiley.
- Mason, I. (1982). A model for assessment of weather forecasts. *Australian Meteorological Magazine*, **30**, 291–303.
- Morris, P. A. (1974). Decision analysis expert use. *Management Science*, **20**, 1233–1241.
- Morris, P. A. (1977). Combining expert judgment: A Bayesian approach. *Management Science* **23**, 679–693.
- Morris, P. M. (1982, January). *The evaluation of subjective probability assessments*. (Working Paper 82-10). College of Business Administration, Northeastern University.
- Murphy, A. H. (1966). A note on the utility of probabilistic predictions and the probability score in the cost-loss ratio decision situation. *Journal of Applied Meteorology*, **5**, 534–537.
- Murphy, A. H. (1969a). On expected-utility measures in cost-loss ratio decision situations. *Journal of Applied Meteorology*, **8**, 989–991.
- Murphy, A. H. (1969b). Measures of the utility of probabilistic predictions in cost-loss ratio decision situations in which knowledge of the cost-loss ratios is incomplete. *Journal of Applied Meteorology*, **8**, 863–873.
- Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, **12**, 595–600.
- Murphy, A. H. (1976). Decision-making models in the cost-loss ratio situation and measures of the value of probability forecasts. *Monthly Weather Review*, **104**, 1058–1065.
- Murphy, A. H. (1977). The value of climatological, categorical and probabilistic forecasts in the cost-loss situation. *Monthly Weather Review*, **105**, 803–816.
- Murphy, A. H., & Epstein, E. S. (1967). Verification of probabilistic predictions: A brief review. *Journal of Applied Meteorology*, **6**, 748–755.
- Murphy, A. H., & Winkler, R. L. (1970). Scoring rules in probability assessment and evaluation. *Acta Psychologica*, **34**, 273–286.
- Murphy, A. H., & Winkler, R. L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society(c)*, **26**, 41–47.
- Sanders, F. (1963). On subjective probability forecasting. *Journal of Applied Meteorology*, **2**, 191–201.

- Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, **66**, 783–801.
- Shuford, E. H., Jr., Albert, A., & Massengill, H. E. (1966). Admissible probability measurement procedures. *Psychometrika*, **31**, 125–145.
- Stiglitz, J. (1974). Risk sharing and incentives in sharecropping. *Review of Economic Studies*, **61**(2), 219–256.
- Swets, J. A., & Pickett, R. M. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory*. New York: Academic Press.
- Thompson, J. C. (1952). On the operational deficiencies in categorical weather forecasts. *Bulletin of the American Meteorological Society*, **33**, 223–226.
- Thompson, J. C., & Brier, G. W. (1955). The economic utility of weather forecasts. *Monthly Weather Review*, **83**, 249–254.
- von Neumann, Jr., & Morgenstern, O. (1953). *Theory of games and economic behavior*. Princeton, NJ: Princeton University.
- Winkler, R. L. (1969). Scoring rules and the evaluation of probability assessors. *Journal of the American Statistical Association*, 1073–1078.
- Winkler, R. L., & Murphy, A. H. (1968). “Good” probability assessors. *Journal of Applied Meteorology*, **2**, 751–758.
- Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance*, **30**, 132–156.
- Yates, J. F. (1984). Evaluating and analyzing probabilistic forecasts. *The UMAP Journal*, **5**, 76–118.

RECEIVED: August 29, 1983