Measures of the Utility of Probabilistic Predictions in Cost-Loss Ratio Decision Situations in which Knowledge of the Cost-Loss Ratios is Incomplete

ALLAN H. MURPHY[1]

The Travelers Research Corporation, Hartford, Conn.

(Manuscript received 9 May 1969, in revised form 21 August 1969)

ABSTRACT

Comparative operational evaluation of probabilistic prediction procedures in cost-loss ratio decision situations in which the evaluator's knowledge of the cost-loss ratio is expressed in probabilistic terms is considered. First, the cost-loss ratio decision situation is described in a utility framework and, then, measures of the expected-utility of probabilistic predictions are formulated. Second, a class of expected-utility measures, the beta measures, in which the evaluator's knowledge of the cost-loss ratio is expressed in terms of a beta distribution, are described. Third, the beta measures are utilized to compare two prediction procedures on the basis of a small sample of predictions. The results indicate the importance, for comparative operational evaluation, of utilizing measures which provide a suitable description of the evaluator's knowledge. In particular, the use of the probability score, a measure equivalent to the *uniform* measure (which is a special beta measure), in decision situations in which the uniform distribution does not provide a suitable description of the evaluator's knowledge, may yield misleading results. Finally, the results are placed in proper perspective by describing several possible extensions to this study and by indicating the importance of undertaking such studies in actual operational situations.

1. Introduction

We are, in this paper, concerned with comparative operational evaluation in cost-loss ratio decision situations. For the purposes of this discussion, comparative operational evaluation is considered to be the evaluation of two forecasters (or prediction procedures), A and B say, in terms of the difference in operational value, or "utility," of their probabilistic predictions. The cost-loss ratio situation is, of course, the familiar two-action, two-state decision situation (refer to Section 2). We are, in particular, concerned with comparative operational evaluation in cost-loss ratio situations in which the evaluator's[2] knowledge of the cost-loss ratios is incomplete (an evaluator's knowledge of the cost-loss ratio is seldom, if ever, complete). We assume that the evaluator expresses his knowledge of the cost-loss ratio in probabilistic terms (complete knowledge is then simply a special case of incomplete knowledge).

Murphy (1969) has shown that, in the two-state situation, whether the evaluator's knowledge is complete or incomplete, if the validity of A's prediction is greater than the validity of B's prediction, then the expected-utility[3] of A's prediction is at least as great as the expected-utility of B's prediction.[4,5] Thus, in the cost-loss ratio situation, a validity measure, e.g., the probability score (Brier, 1950), represents an ordinal measure on the expected-utilities, i.e., such a measure determines the sign of the expected-utility difference. However, we would, of course, prefer an interval, or exact, measure. That is, we would like to be able to determine the magnitude (as well as the sign) of the expected-utility difference.

Murphy (1966) has also shown that the expected-utility of a probabilistic prediction, in a cost-loss ratio

[1] Present affiliation: Department of Meteorology and Oceanography, The University of Michigan, Ann Arbor.

[2] Evaluators are individuals concerned with assessing the "goodness" of predictions (or prediction procedures). In particular, an evaluator may, or may not, be a meteorologist or a decision maker. However, in this paper, we assume that the evaluator of concern possesses the same knowledge as the meteorologist and the decision maker concerning the probabilities and the utilities, respectively.

[3] In this paper the term "expected-utility" refers to both the (first-order) expectation of the utility of a prediction (or an action) over a (probability) distribution on the states, and the (second-order) expectation of the utility of a prediction over a distribution on the cost-loss ratio.

[4] Winkler and Murphy (1968) have indicated that probabilistic predictions should possess both *normative* and *substantive* standards of "goodness." We are concerned, in this paper, only with substantive "goodness." Murphy and Epstein (1967) have shown that, in a meteorological context, substantive evaluation consists, in essence, of empirical ("scientific") evaluation and operational ("economic") evaluation. *Validity* is a "desirable" attribute (of a prediction) from the point of view of empirical evaluation, while *utility* is a "desirable" attribute from the point of view of operational evaluation.

[5] This statement properly refers only to comparative evaluation for individual predictions. For necessary and sufficient conditions for a similar result to hold for average expected-utilities for a collection of predictions, refer to Murphy (1969). For the purposes of this paper we simply observe that, for a collection of predictions, the fact that A's average validity is greater than B's average validity is not sufficient to guarantee that A's average expected-utility is greater than (or equal to) B's average expected-utility.

situation in which the evaluator's knowledge of the cost-loss ratio is expressed in terms of a *uniform* (probability) distribution, is linearly related to the validity of the prediction as measured by the probability score. Thus, in situations in which the evaluator's knowledge of the cost-loss ratio is incomplete and is suitably described by a uniform distribution (a *special* case), the probability score, or an *equivalent* measure,[6] is an appropriate measure for operational as well as empirical evaluation. Specifically, for comparative evaluation in the uniform case, the difference between the expected-utilities of two predictions is directly proportional to the difference between their probability scores (validities).

However, if the evaluator's knowledge of the cost-loss ratio is such that the uniform distribution does not provide a suitable description (the *general* case), then the expected-utility of a prediction is not a linear function of the probability score (Murphy, 1969). Thus, the probability score, which is an appropriate measure of expected-utility (as well as validity) in the uniform case, is not an appropriate measure of expected-utility in the non-uniform case. In this paper we describe some results which, at least in part, answer certain questions related to comparative evaluation in the non-uniform case. For example: 1) What measures of expected-utility are appropriate when the evaluator's knowledge consists of a non-uniform distribution? 2) What are the differences between expected-utilities in the uniform and non-uniform cases both for individual predictions and for collections of predictions? 3) To what extent is the probability score an appropriate measure for operational evaluation in the non-uniform case? In Section 2 we describe, in brief, the cost-loss ratio decision situation in a utility framework. In Section 3 we define, and briefly describe the properties and behavior of, the expected-utility measures. A particular class of expected-utility measures, the *beta* measures, is described in Section 4. These measures are based upon the assumption that the evaluator's knowledge can, in many situations, be suitably expressed in the form of a beta distribution. In Section 5 we describe the results of the comparative evaluation of two prediction procedures when the evaluator's knowledge is expressed in terms of different members of the class of beta measures. Section 6 contains a brief conclusion and summary.

TABLE 1a. The cost-loss ratio decision situation in a utility framework.

Action	State: Adverse Weather (W)	State: No Adverse Weather (NW)	
Protect (P)	X	X	Utilities
Do Not Protect (DNP)	0	1	Utilities
	p_W	p_{NW}	Probabilities

TABLE 1b. The expected-utility decision rule in the decision situation depicted in Table 1a.

P	if	$p_{NW}<X$
P or DNP	if	$p_{NW}=X$
DNP	if	$p_{NW}>X$

2. Cost-loss ratio decision situation

The cost-loss ratio decision situation (Thompson, 1952) is depicted in a utility framework in Table 1a [refer to Murphy (1966)]. The elements of such a decision situation of particular interest, from the point of view of (quantitative) evaluation, are the *probabilities* and the *utilities*. We assume, in this framework, that the evaluator's knowledge[7] of the states (of the atmosphere) is expressed in terms of a probability vector $\mathbf{p}=(p_W,p_{NW})$, where $p_W(p_{NW})$ is the probability of (no) adverse weather $(p_W,p_{NW}\geq 0, p_W+p_{NW}=1)$. The elements of the utility matrix $\mathbf{U}$, i.e., the members of the set $\{X,X,0,1\}$, are *standard* utilities (in utiles) that express the decision maker's preferences for the respective consequences. The utility X, a number between zero and one, represents the cost-loss ratio in this framework.[8] We also assume, in this framework, that the appropriate decision criterion is the expected-utility decision criterion. Thus, the decision maker selects the action (P or DNP) which maximizes his expected-utility. The decision rule in this situation is described in Table 1b.

3. Expected-utility measures

a. Knowledge of the probabilities and the utilities

In this paper we assume that the meteorologist's knowledge of the probabilities (on the states) is complete and that the decision maker's knowledge of the utility X is, in general, incomplete. (Thus, the evaluator's knowledge of the probabilities and utilities is complete and incomplete, respectively; refer to Footnote 2.) Further, we assume that the decision maker expresses his knowledge of X in probabilistic terms, i.e., we assume that the decision maker's knowledge of the utility X is expressed in terms of a (probability) density function $f(x)$. Thus, the "utility" of a prediction is its *expected-utility* (with the expectation being taken with respect to both the probability distribution on the states and the distribution of the utility X; refer to Footnote 3). Note that the expected-utility of a prediction depends upon the nature of the density function $f(x)$ as well as the probabilities, the utility X, and the observed state (see below).

[6] Two evaluation measures are equivalent if the measures are linearly related (Murphy, 1969).

[7] The evaluator's knowledge is assumed to be identical to that of the meteorologist (refer to Footnote 2).

[8] Specifically, X equals one minus the cost-loss ratio.

b. Expected-utility measure $E_{\mathbf{p}}[U(X)]$

Let $E_{\mathbf{p}}[U(X)]$, or simply $U(X)$, denote the expected-utility of a prediction $\mathbf{p}$ (the expectation being taken with respect to the "distribution" of the probabilities). For a particular value of the utility X, x (say), $U(X)$ becomes $U(x)$, where

$$U(x)=xd(x,p_{NW})+[\tfrac{1}{2}x\delta_W+\tfrac{1}{2}(x+1)\delta_{NW}]e(p_{NW},x)+\delta_{NW}d(p_{NW},x),$$

or

$$U(x)=xd(x,p_{NW})+\tfrac{1}{2}(x+\delta_{NW})e(p_{NW},x)+\delta_{NW}d(p_{NW},x), \quad (1)$$

in which

$$d(r,s)=\begin{cases}1, & \text{if } r>s\\ 0, & \text{if } r\leq s\end{cases}$$

$$e(r,s)=\begin{cases}1, & \text{if } r=s\\ 0, & \text{if } r\neq s\end{cases}$$

$$\delta_W(\delta_{NW})=\begin{cases}1(0), & \text{if } W \text{ obtains}\\ 0(1), & \text{if } NW \text{ obtains}\end{cases}$$

The function $d(r,s)$ is a step function (with a single step at $r=s$); the function $e(r,s)$ is the "unit" function (the integral of the unit function over any interval is zero); and $\delta_W(\delta_{NW})$ is the (Kronecker) delta function. The range of $U(x)$, in (1), is the closed unit interval $[0,1]$. However, $U(x)$ can assume only five different values,[9] namely, 0, $\frac{1}{2}x$, x, $\frac{1}{2}(1+x)$ and 1. $U(x)$, a step-function with a single step at $p_{NW}=x$, is depicted in Fig. 1.

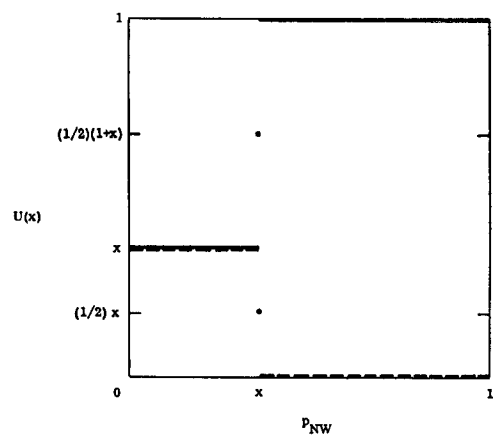

FIG. 1. The expected-utility of a prediction $\mathbf{p}$, $U(x)$, as a function of p_{NW}. The solid (dashed) line represents $U(x)$ when δ_{NW} equals one (zero).

[9] The value of $U(x)$ when $p_{NW}=x$ is, to a certain extent, arbitrary. That is, in general, $U(x)=a_1x$ when $\delta_W=1$ and $U(x)=a_1x+a_2$ when $\delta_{NW}=1$, where a_1, $a_2\geq 0$ and $a_1+a_2=1$. We have chosen a_1 and a_2 to be equal to one-half.

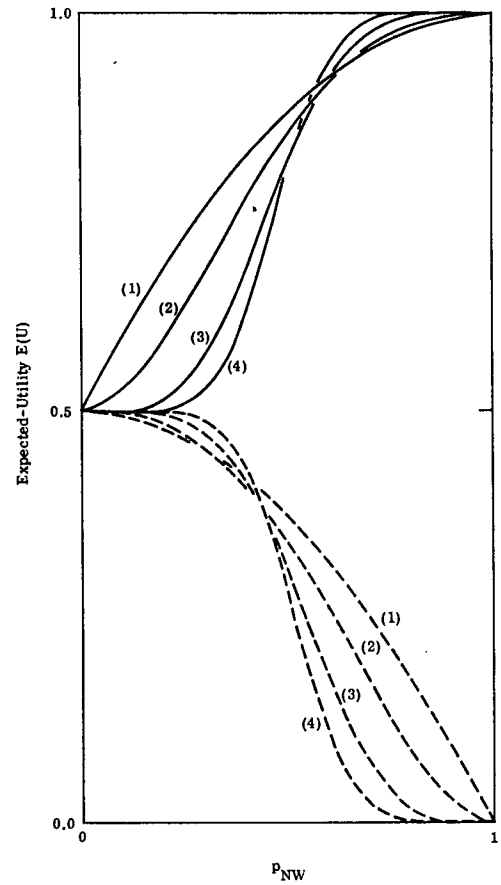

FIG. 2a. The expected-utility of a prediction $\mathbf{p}$, $E(U)$, as a function of p_{NW} when $X\sim\beta(a,b)$, where $E(X)=\frac{1}{2}$: (1) $a=b=1$; (2) $a=b=2$; (3) $a=b=5$; (4) $a=b=10$. Solid (dashed) curves define $E(U)$ when δ_{NW} equals one (zero).

The average utility of a collection of K predictions $\mathbf{p}_k$ (for the same decision situation) is $\bar{U}(x)$, where

$$\bar{U}(x)=(1/K)\sum_k U_k(x), \quad (k=1,\cdots,K). \quad (2)$$

c. Expected-utility measure $E_X\{E_{\mathbf{p}}[U(X)]\}$

The expected-utility of a prediction $\mathbf{p}$, when the expectation is taken, in addition, with respect to the distribution of the utility X, is $E_X\{E_{\mathbf{p}}[U(X)]\}$, or simply $E(U)$, where

$$E(U)=\int_0^1 U(x)f(x)dx,$$

or, from (1),

$$E(U)=\int_0^1 [xd(x,p_{NW})+\tfrac{1}{2}(x+\delta_{NW})e(p_{NW},x)+\delta_{NW}d(p_{NW},x)]f(x)dx,$$

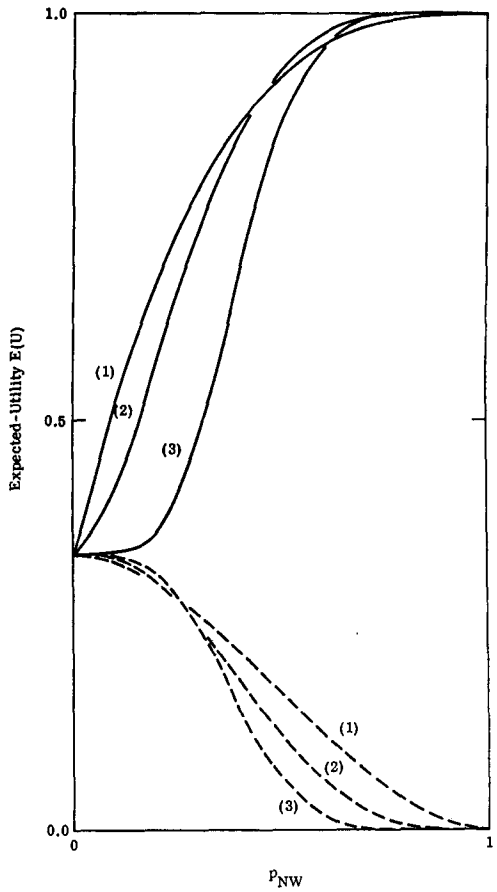

FIG. 2b. The expected-utility of a prediction $\mathbf{p}$, $E(U)$, as a function of p_{NW} when $X \sim \beta(a,b)$, where $E(X)=\frac{1}{2}$: (1) $a=1$, $b=2$; (2) $a=2$, $b=4$; (3) $a=4$, $b=8$. Solid (dashed) curves define $E(U)$ when δ_{NW} equals one (zero).

or

$$E(U)=\int_{p_{NW}}^{1} xf(x)dx+\delta_{NW}\int_{0}^{p_{NW}} f(x)dx. \quad (3)$$

Note that $E(U)$, in (3), depends upon the nature of the evaluator's (decision maker's) knowledge, i.e., upon the form of the density function $f(x)$. [For an examination of the behavior of $E(U)$ for specific density functions refer to Section 4.] $E(U)$ is a continuous function of p_{NW} (or p_W, since $p_W=1-p_{NW}$). The range of $E(U)$ is the closed unit interval $[0,1]$. $E(U)$ equals zero if p_{NW} equals one and δ_{NW} equals zero, while $E(U)$ equals one if p_{NW} equals one and δ_{NW} equals one.

The average expected-utility of a collection of K predictions $\mathbf{p}_k$ (for the same decision situation) is $\overline{E(U)}$, where

$$\overline{E(U)}=(1/K)\sum_{k} E(U_k), \quad (k=1,\cdots,K). \quad (4)$$

4. Beta measures

a. Beta distribution

As indicated in Section 3, the evaluator's knowledge of the utility X is expressed in terms of a density function $f(x)$ and, in order to determine the expected-utility $E(U)$, in (3), we must specify $f(x)$. We would like to select a *family* of density functions $f(x)$ which is "rich," i.e., a family which is able to suitably represent the evaluator's knowledge in different decision situations. The beta distribution, a distribution which can assume a variety of different shapes for different values of its parameters, is such a distribution.

The beta density function for a random variable X is $f(x)$, where

$$f(x)=\frac{(a+b-1)!}{(a-1)!(b-1)!}x^{a-1}(1-x)^{b-1},$$
$$(0\leq x\leq 1;\ a,\ b>-1), \quad (5)$$

in which a and b are the parameters of the distribution. The expected value of X, $E(X)$, is $a/(a+b)$, while the variance of X, $V(X)$, is $ab/[(a+b)^2(a+b+1)]$. In this paper we shall restrict the values of the parameters a and b to positive integer values. With such a restriction the density function $f(x)$, in (5), is unimodal. If a and b

TABLE 2. The sample of predictions prepared by procedures A and B. The frequency of predictions f_p and observations f_0, as well as the relative frequency f_0/f_p, are indicated for different values of p_{NW}.

	Procedure A			Procedure B		
p_{NW}	Prediction frequency f_p	Observation frequency f_0	Relative frequency f_0/f_p	Prediction frequency f_p	Observation frequency f_0	Relative frequency f_0/f_p
0.0						
0.1						
0.2	4	1	0.250	4	0	0.000
	(3)*	(1)	(0.333)	(3)	(0)	(0.000)
0.3	1	0	0.000	2	0	0.000
	(0)	(0)	(——)	(1)	(0)	(0.000)
0.4	11	3	0.273	6	2	0.333
	(10)	(3)	(0.300)	(5)	(2)	(0.400)
0.5	26	9	0.346	16	7	0.433
	(24)	(9)	(0.375)	(14)	(7)	(0.500)
0.6	48	30	0.625	31	18	0.581
	(39)	(25)	(0.641)	(22)	(13)	(0.591)
0.7	48	35	0.729	57	36	0.632
	(38)	(27)	(0.711)	(47)	(28)	(0.596)
0.8	50	35	0.700	112	91	0.813
	(36)	(24)	(0.667)	(98)	(80)	(0.816)
0.9	70	61	0.871	233	215	0.923
	(34)	(28)	(0.824)	(197)	(182)	(0.924)
1.0	300	291	0.970	97	96	0.990
	(211)	(202)	(0.957)	(8)	(7)	(0.875)
Total or average	558	465	0.833	558	465	0.833
	(395)	(319)	(0.808)	(395)	(319)	(0.808)

* The numbers in parentheses refer to the "distinct" (sub-) collection.

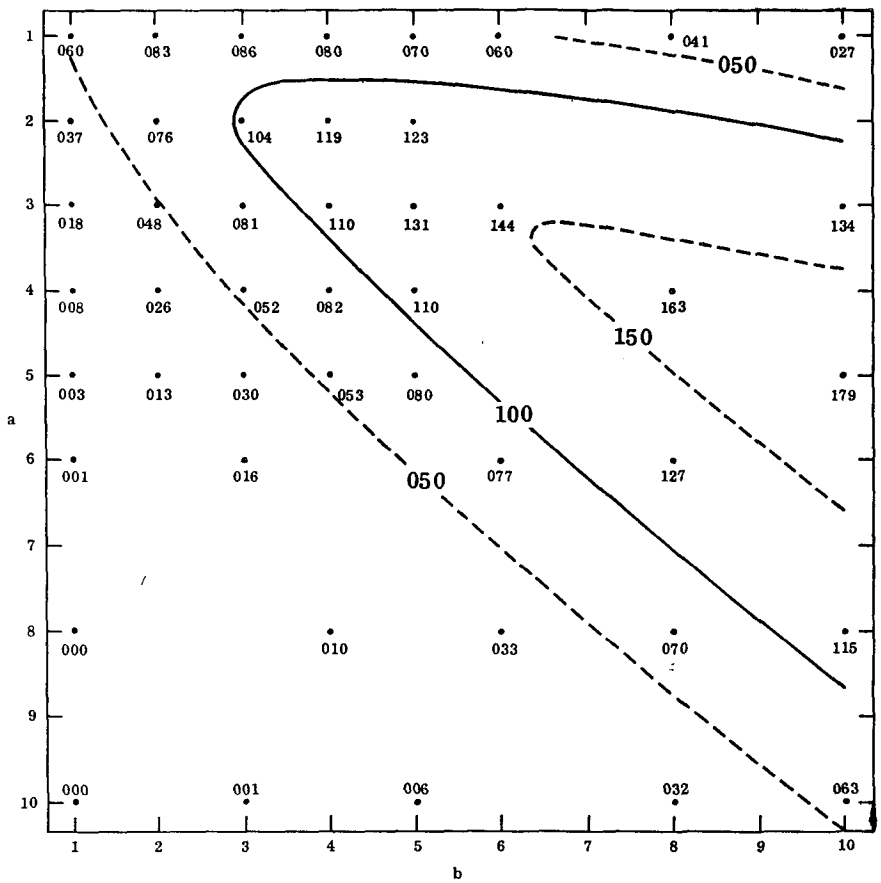

FIG. 3a. The expected-utility difference $E(U)D\times10^{-3}$ as a function of the values of the parameter set $\{a,b\}$ when $p_W^A=0.8$, $p_W^B=0.6$ and $\delta_W=1$.

are equal, $f(x)$ is symmetric, while if a is greater (less) than b, $f(x)$ is skewed to the left (right). When a and b are both equal to one the beta distribution reduces to the uniform distribution.

b. Beta measures

We consider both the general (non-uniform) case and the special (uniform) case.

1) General case. The expected-utility measure $E(U)$, when the cost-loss ratio X has a beta distribution, is, from (3) and (5),

$$E(U)=\frac{(a+b-1)!}{(a-1)!(b-1)!}\left[\int_{p_{NW}}^{1}x^{a}(1-x)^{b-1}dx+\delta_{NW}\int_{0}^{p_{NW}}x^{a-1}(1-x)^{b-1}dx\right],$$

or, since a and b are positive integers,

$$E(U)=\frac{a}{a+b}\left[1-\sum_{j=a+1}^{a+b}\frac{(a+b)!}{j!(a+b-j)!}p_{NW}^{j}(1-p_{NW})^{a+b-j}\right]+\delta_{NW}\sum_{j=a}^{a+b-1}\frac{(a+b-1)!}{j!(a+b-1-j)!}\times p_{NW}^{j}(1-p_{NW})^{a+b-1-j}, \quad (6)$$

[refer to Raiffa and Schlaifer (1961), pp. 216–218]. $E(U)$, in (6), is depicted in Figs. 2a and 2b for selected values of the parameters a and b. Note, in Fig. 2a, that as $c(=a=b)$ increases when $E(X)=\frac{1}{2}$, the expected-utility measure $E(U)$ approaches the utility measure $U(\frac{1}{2})$, a step-function with a single step at $p_{NW}=\frac{1}{2}$. Further, note, in Fig. 2b, that as $c(=2a=b)$ increases when $E(X)=\frac{1}{3}$, $E(U)$ approaches $U(\frac{1}{3})$, a step-function with a step at $p_{NW}=\frac{1}{3}$.

2) Special case. A special case of particular interest arises when the parameters a and b are both equal to one, in which case the beta distribution, in (5), reduces to the uniform distribution. In this case the expected-utility measure $E(U)$, in (6), becomes

$$E(U)=\tfrac{1}{2}(1-p_{NW}^{2})+\delta_{NW}p_{NW},$$

or

$$E(U)=-\tfrac{1}{4}PS+\tfrac{1}{2}(1+\delta_{NW}), \quad (7)$$

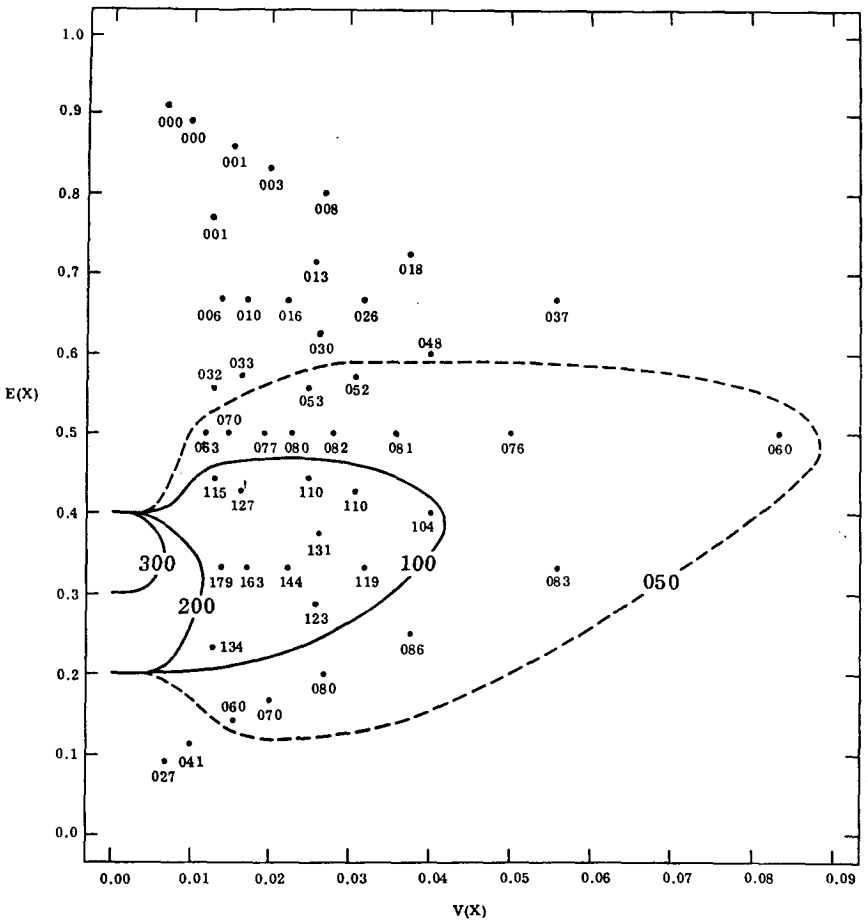

FIG. 3b. The expected-utility difference $E(U)D\times10^{-3}$ as a function of the expected value $E(X)$ and the variance $V(X)$ when $p_W^A=0.8$, $p_W^B=0.6$ and $\delta_W=1$.

where PS is the probability score. Thus, in the uniform case the expected-utility of a prediction is linearly related to the probability score for that prediction. Eq. (7) was obtained, independently, by Hunt (1963) and Murphy (1965, 1966). $E(U)$, in the uniform case, is depicted in Fig. 2a.

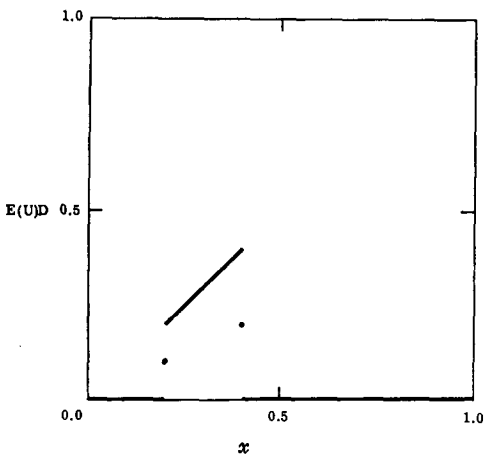

FIG. 3c. The expected-utility difference $E(U)D$ when $V(X)$ equals zero as a function of x when $p_W^A=0.8$, $p_W^B=0.6$ and $\delta_W=1$.

5. Comparative evaluation: Beta measures

In this section we utilize different beta measures to compare two prediction procedures on the basis of a sample of predictions. The results are presented to illustrate the effect of utilizing different beta measures upon comparative evaluation. First we briefly describe the sample of predictions and introduce some additional notation.

a. Sample of predictions

The sample consists of probabilistic predictions prepared by two procedures, A and B say, on 558 occasions. The nature of the procedures themselves is not of particular importance.[10] The probabilities which constitute

[10] Procedure A is a "subjective" procedure while procedure B is an "objective" procedure. A's predictions were prepared utilizing B's predictions as "objective" aids.

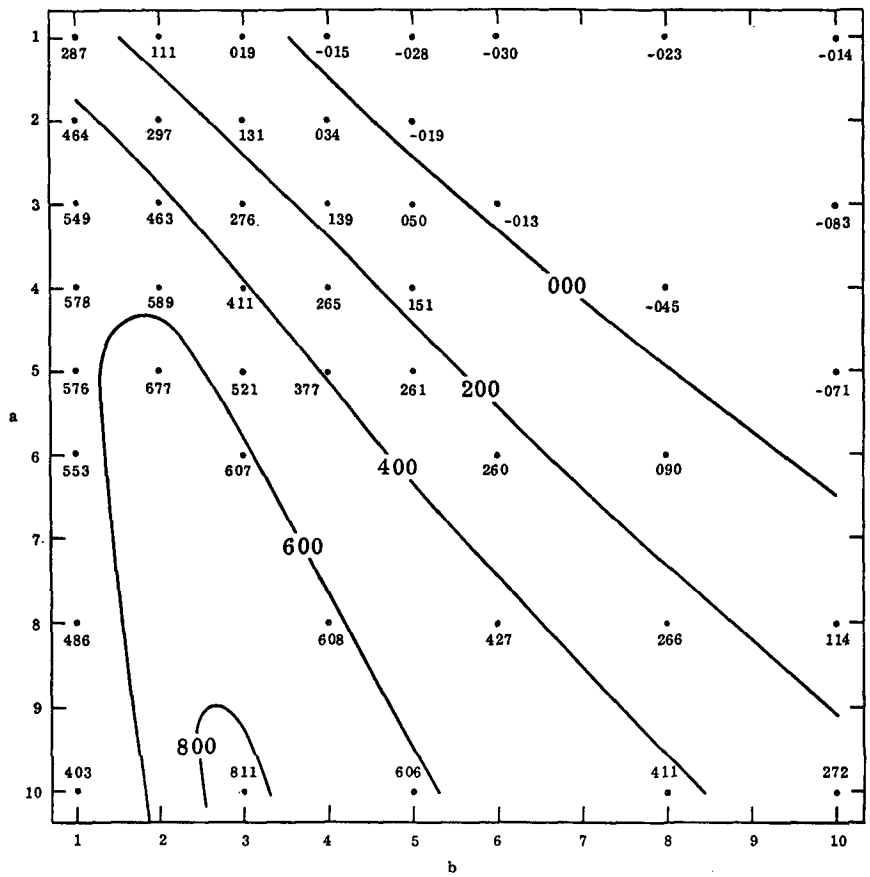

FIG. 4a. The average expected-utility difference $\overline{E(U)D}\times 10^{-5}$ as a function of the parameter set $\{a,b\}$ for the complete collection of predictions.

the predictions prepared by the procedures were expressed in tenths [i.e., $p_W=0.0(0.1)1.0$]. Thus, the number of different predictions possible is only eleven and, not unexpectedly, the predictions $\mathbf{p}^A$ and $\mathbf{p}^B$ were identical on 163 occasions. The sample of predictions is summarized in Table 2 for both the complete collection (558 predictions) and the "distinct" (sub-) collection (395 predictions).

b. Notation

Let $E^A(U)$ and $E^B(U)$ denote the expected-utility of predictions $\mathbf{p}^A$ and $\mathbf{p}^B$ prepared by procedures A and B, respectively. Let $E(U)D$ denote the difference between $E^A(U)$ and $E^B(U)$, i.e., $E(U)D=E^A(U)-E^B(U)$. Further, let the average expected-utilities and the average expected-utility difference be denoted by $\overline{E^A(U)}$, $\overline{E^B(U)}$ and $\overline{E(U)D}$, respectively. Finally, let $PSD(\overline{PSD})$ denote the respective difference in the (average) probability scores.

c. Individual prediction

Consider occasion 437 on which $\mathbf{p}^A=(0.8, 0.2)$, $\mathbf{p}^B=(0.6, 0.4)$ and $\delta_W=1$.

The expected-utility difference $E(U)D$ on this occasion is depicted in Fig. 3a as a function of the values of the parameter set $\{a,b\}$. Note that $E(U)D$ for $\{1,1\}$, i.e., in the *uniform* case, is 0.060. Thus, $E(U)D$ is nonnegative for all sets $\{a,b\}$. The values of $E(U)D$ are small(est) for distributions (of X) for which $E(X)$ is large (near one) or small (near zero), while the values of $E(U)D$ are large(st) for distributions for which $E(X)$ is between 0.2 and 0.4.[11] The latter (as well as the former) result is reasonable since, when $0.2<E(X)<0.4$, the likelihood that the predictions $\mathbf{p}^A$ and $\mathbf{p}^B$ will lead to different actions is large (and, since $p_W^A>p_W^B$ and $\delta_W=1$, $\mathbf{p}^A$ will lead to a "better" action than $\mathbf{p}^B$).

In order to facilitate an examination of the effect of the evaluator's knowledge of X on evaluation, the expected-utility difference $E(U)D$ is depicted in Fig. 3b as a function of the expected value $E(X)$ and the variance $V(X)$. The "critical region" for evaluation, i.e., the region in which a small change in the value of p_{NW} (or p_W) can lead to a large change in the value of $E(U)$, is located in the vicinity of $E(X)$ (refer to Figs. 2a and

[11] Since $f(x)$ is unimodal, $E(X)$ represents a good indicator of the location of the (mass of the) distribution.

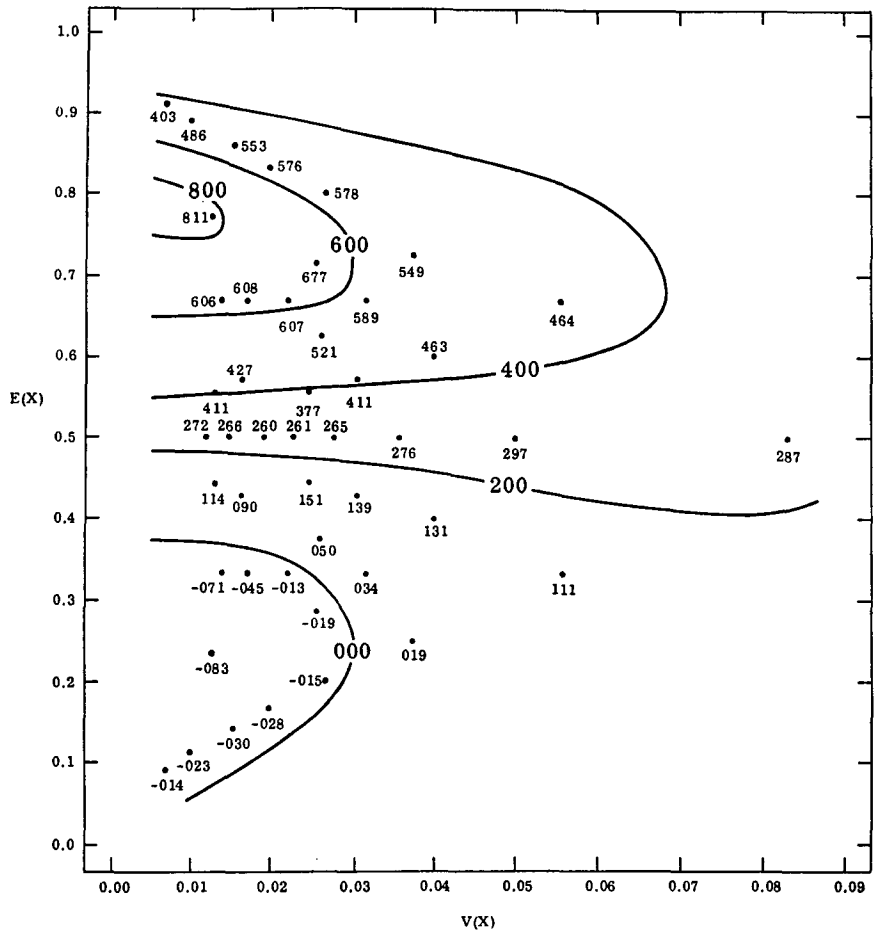

FIG. 4b. The average expected-utility difference $\overline{E(U)D}\times 10^{-5}$ as a function of $E(X)$ and $V(X)$ for the complete collection of predictions.

2b). The presence or absence of a critical region in comparative evaluation depends, as indicated in Fig. 3b, on the relative location of the "points" p^A_{NW}, p^B_{NW} and $E(X)$. If $E(X)$ is much less than, or much greater than, both p^A_{NW} and p^B_{NW}, then a critical region is not present. However, if $E(X)$ is located between p^A_{NW} and p^B_{NW}, then a critical region is present in the vicinity of $E(X)$. The "intensity" of such a critical region increases as the variance $V(X)$ decreases (i.e., as the evaluator's knowledge of X increases). Note that when the variance $V(X)=0$, i.e., when the evaluator's knowledge is complete, $E(X)=x[=a/(a+b)]$. If $0.2<x<0.4$, $E(U)D=x$, otherwise $E(U)D=0$. $E(U)$,D on this occasion, is depicted in Fig. 3c as a function of x.

The variability of $E(U)D$ as a function of the values of the parameter set $\{a,b\}$ is also of interest from the point of view of the use of the probability score as a measure of the "utility" (as well as the validity) of individual predictions in the non-uniform, as well as the uniform, case (refer to Fig. 3a). Note that $E(U)D$ in the uniform case is, as indicated, 0.060 [and, since $E(U)D=-\frac{1}{4}PSD$, $PSD=-0.240$]. However, $E(U)D$ is 0.179 for $\{10,5\}$, while $E(U)D$ is 0.000 for $\{1,10\}$. Thus, if A and B are compared on the basis of the probability score when the evaluator's knowledge of X is such that $X\sim\beta(10,5)$, then $E(U)D$ will be seriously underestimated. On the other hand, if $X\sim\beta(1,10)$, then $E(U)D$ is overestimated. Thus, the importance of the evaluator expressing his knowledge of the cost-loss ratio in a suitable form and then utilizing the appropriate expected-utility measure is evident.

d. Collections of predictions

The average expected-utility difference $E(U)D$ for the complete collection of predictions and the "distinct" (sub-) collection of predictions are depicted in Figs. 4a and 5a, respectively, as a function of the values of the parameter set $\{a,b\}$. Figs. 4b and 5b depict the average expected-utility difference $E(U)D$ for the complete collection and the "distinct" collection, respectively, as a function of the expected value $E(X)$ and the variance $V(X)$. The results for the "distinct" collection are presented separately since the evaluator may want to base

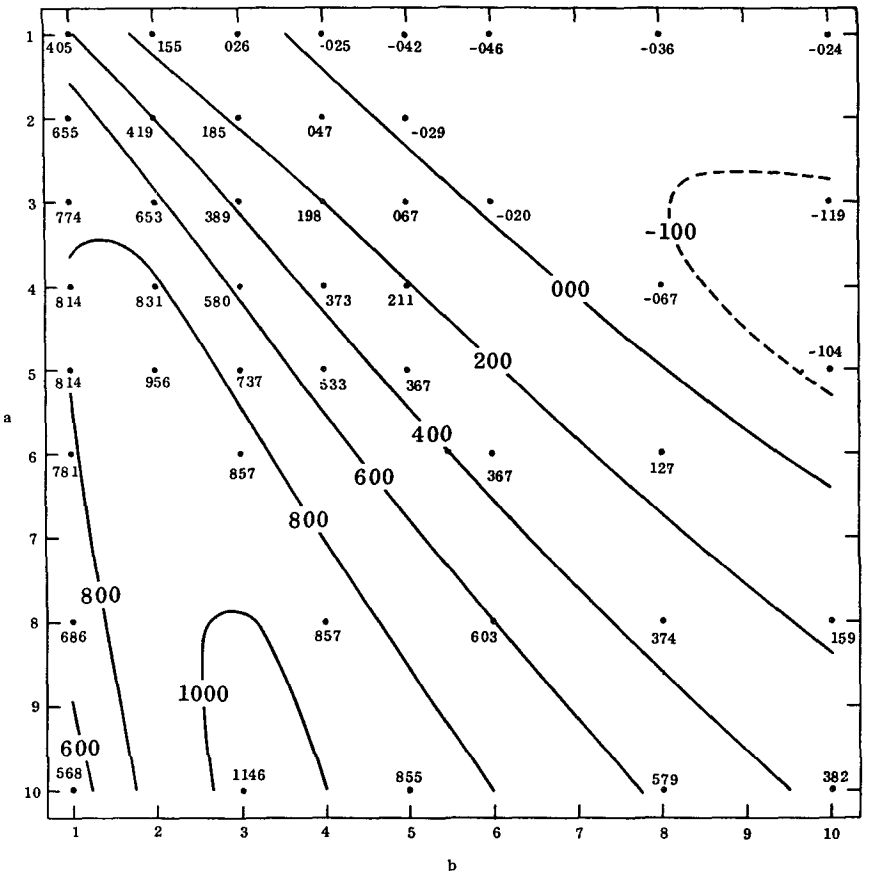

FIG. 5a. Same as Fig. 4a except for the "distinct" collection of predictions.

the comparative evaluation on the collection of occasions on which the predictions are different.

1) Complete collection. Note, in Fig. 4a, that $E(U)D$ for $\{1,1\}$ is 0.00287 [and, since $E(U)D=-\frac{1}{4}PSD$, $PSD=-0.01148$]. However, the fact that $E(U)D$ is positive for $\{1,1\}$ does not guarantee that $E(U)D$ will be positive for all values of the parameter set $\{a,b\}$ (refer to Footnote 5). Note that $E(U)D$ becomes negative when a is small and b is large, while $E(U)D$ increases when a is large and b is small. In other words, $E(U)D$ becomes negative when $E(X)$ decreases, while $E(U)D$ increases when $E(X)$ increases. Thus, if procedures A and B are compared on the basis of their average probability scores when the evaluator's knowledge of X is such that $X\sim\beta(3,10)$ (for example), then $\overline{E(U)D}$ will indicate that A is "better" than B when, in fact, B is "better" than A. On the other hand, if $X\sim\beta(10,3)$, $\overline{E(U)D}$ will be seriously underestimated.

The variability of the average expected-utility difference $\overline{E(U)D}$ is depicted in Fig. 5a as a function of $E(X)$ and $V(X)$. The critical regions for a collection of predictions will, in general, be less "intense" than the critical regions for individual predictions simply because the differences are *average* differences and, as such, represent a collection of occasions 1) on some of which A is "better" than B and on some of which B is "better" than A (in general), and 2) on which different probability values appear in the predictions (in general). The effects of such factors are particularly evident in Fig. 5a (as well as Fig. 4a) for the rather small and similar samples of predictions considered in this paper. However, the region in Fig. 5a corresponding to the maximum positive differences in average expected-utility $\overline{E(U)D}$ is in the vicinity of $E(X)=0.8$ and, from Table 2, $\overline{p}^A_{NW}=0.867$ and $\overline{p}^B_{NW}=0.836$. Thus, as we might expect, the largest values of $\overline{E(U)D}$ are in the region in which a relatively large number of predictions in the two samples are "separated" by the value of $E(X)$.

The location of the isopleths of equal average expected-utility $E(U)D$ in Fig. 5a (in particular) should be considered to be approximate, particularly when $V(X)$ approaches zero, since discontinuities in $E(U)D$ occur along the $E(X)$ axis.[12] However, since the samples did not contain any predictions for which $p_{NW}<0.2$ (refer to Table 2), $E(U)D=0$ when $V(X)=0$ and $E(X)$ $(=x)<0.2$. The determination of the value of $E(U)D$

[12] The discontinuities occur at the points $E(X)=0.2(0.1)1.0$ for the sample of predictions.

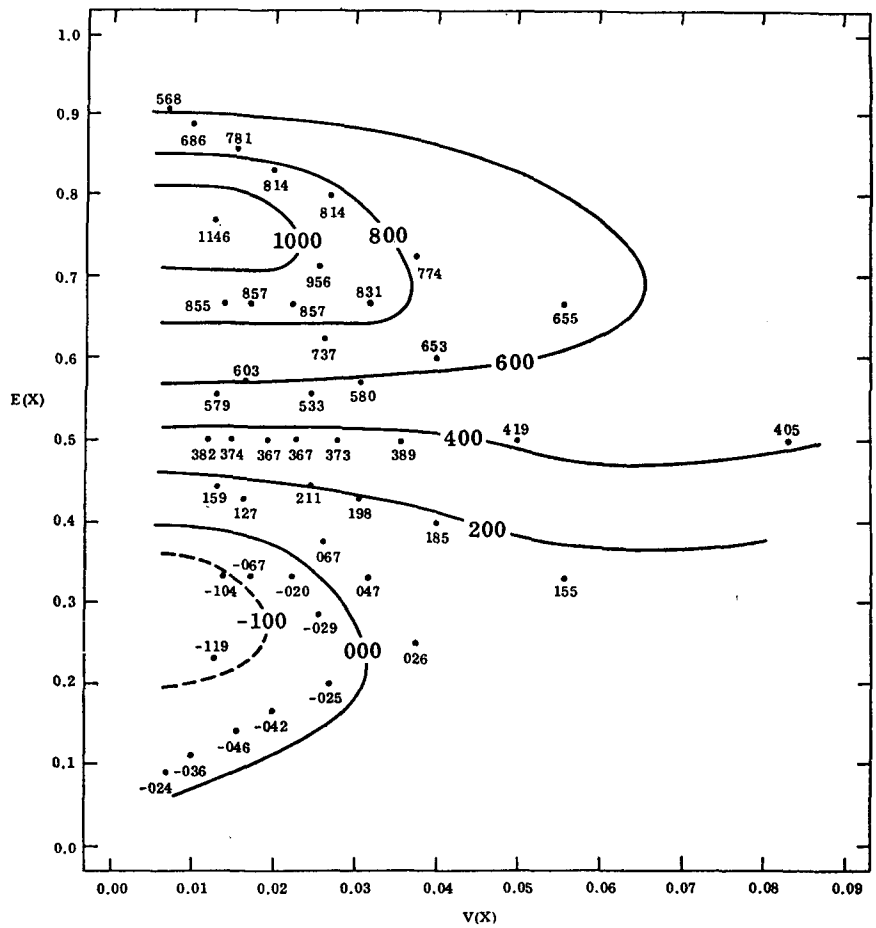

FIG. 5b. Same as Fig. 4b except for the "distinct" collection of predictions.

when $V(X)=0$ for $E(X)(=x)>0.2$ requires consideration of the individual pairs of A's and B's predictions.

2) "Distinct" (sub-) collection. The results for the "distinct" collection, as depicted in Figs. 4b and 5b, are similar to the results for the complete collection (as depicted in Figs. 4a and 5a, respectively). However, the values of $\overline{E(U)D}$ for the "distinct" collection are, of course, larger, in absolute value, than the corresponding values of $\overline{E(U)D}$ for the complete collection. The statements concerning the isopleths in Fig. 5a when $V(X)$ is small, as well as the discontinuities when $V(X)$ equals zero, are equally applicable to Fig. 5b.

3) Summary. The variability of $\overline{E(U)D}$ for different values of the parameter set $\{a,b\}$ indicates the importance for the evaluator, and for comparative evaluation, of the selection of the appropriate beta distribution.

6. Conclusions

In this paper we have described a class of expected-utility measures for comparative operational evaluation, the beta measures, which permit the evaluator, in many situations, to express his knowledge of the cost-loss ratio in a suitable form. Further, we have examined, for a small collection of forecasts, the effect of utilizing different beta measures upon the comparative evaluation of two prediction procedures. The results of this preliminary study reveal the importance of the evaluator expressing his knowledge of the cost-loss ratio in terms of the appropriate beta distribution, rather than simply utilizing the probability score (for example), which implies a uniform distribution, for comparative operational evaluation. The results of this study, and the study itself, should, however, be considered as illustrative rather than definitive. Clearly, the exploratory research described in this paper suggests several areas for future research.

First, we could extend the results of this paper to beta distributions in which the parameters need not be integers and, then, we could examine distributions other than the beta distribution. Second, we could extend the results to situations in which the evaluator's knowledge of the probabilities as well as, or instead of, the cost-loss ratio is incomplete. Some relevant results have been ob-

tained in a decision making context by Fishburn (1964, 1965), Fishburn *et al.* (1968) and Fishburn and Murphy (1969). However, only Murphy (1969) has obtained results in such situations in an evaluation context. Third, we could extend such studies to the general two-state, two-action decision situation (Murphy, 1969) and to decision situations in which the number of actions and/or states exceeds two. Finally, and most important from a practical point of view, we could, and must, study the effect of utilizing different distributions to describe the evaluator's knowledge upon the results of comparative evaluation within the context of operational situations such as the construction industry (Russo, 1966). Such studies would reveal the relative importance (or unimportance) of a suitable description of the evaluator's knowledge in actual operational situations.

Acknowledgments. The author would like to acknowledge the valuable comments of Dr. E. S. Epstein of The University of Michigan and Dr. R. L. Winkler of Indiana University on earlier versions of this manuscript.

REFERENCES

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3.

Fishburn, P. C., 1964: *Decision and Value Theory.* New York, Wiley, 451 pp.

——, 1965: Analysis of decisions with incomplete knowledge of probabilities. *Operations Res.*, **13**, 217–237.

——, A. H. Murphy and H. H. Isaacs, 1968: Sensitivity of decisions to probability estimation errors: A re-examination. *Operations Res.*, **16**, 254–267.

——, and ——, 1969: The uncertainty of uncertainty. *Proc. Fourth Intern. Conf. Operational Research*, New York, Wiley, 906–913.

Hunt, J. A., 1963: Decision theory and subjective probability in meteorological forecasts. M. S. thesis, Dept. of Electrical Engineering, Massachusetts Institute of Technology.

Murphy, A. H., 1965: The cost-loss ratio decision situation. Dept. of Meteorology and Oceanography, University of Michigan, unpublished manuscript.

——, 1966: A note on the utility of probabilistic predictions and the probability score in the cost-loss ratio decision situation. *J. Appl. Meteor.*, **5**, 534–537.

——, 1969: The evaluation of probabilistic predictions in meteorology. Ph.D. thesis, Dept. of Meteorology and Oceanography, University of Michigan.

——, and E. S. Epstein, 1967: Verification of probabilistic predictions: A brief review. *J. Appl. Meteor.*, **6**, 748–755.

Raiffa, H., and R. Schlaifer, 1961: *Applied Statistical Decision Theory.* Boston, Harvard University, Graduate School of Business Administration, 356 pp.

Russo, J. A., 1966: The economic impact of weather on the construction industry in the United States. *Bull. Amer. Meteor. Soc.*, **47**, 967–972.

Thompson, J. C., 1952: On the operational deficiencies in categorical weather forecasts. *Bull. Amer. Meteor. Soc.*, **33**, 223–226.

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