Ordinal relationships between measures of the "accuracy" and "value" of probability forecasts: preliminary results[1, 2, 3]

By ALLAN H. MURPHY[4], *Department of Meteorology and Oceanography, University of Michigan, Ann Arbor, Michigan, USA*

(Manuscript received October 15, 1971; revised version July 10, 1972)

ABSTRACT

The preliminary results of an investigation of the relationships, on an ordinal scale, between particular measures of the "accuracy" (the validity measure) and "value" (the utility and expected-utility measures) of probability forecasts are described. For *individual*, comparable forecasts the validity measure represents an ordinal measure on utility and expected utility, i.e., the validity measure is able to order, or rank, forecasters and/or forecast procedures according to the utility or expected utility of their forecasts. Although forecasts, in general, are not often comparable, all forecasts are comparable in two-state, i.e., "weather—no weather", situations and, in addition, these relationships can be expected to hold in most situations in which forecasts are nearly comparable. For *collections* of individually comparable forecasts similar relationships hold between average validity and average utility and expected utility in certain special situations. The nature of the relationships and scope of the results in both general and special situations are indicated by means of sample collections of forecasts.

1. Introduction

Meteorologists concerned with the problem of forecast evaluation have speculated, from time to time, about the existence of relationships between the "accuracy" and "value" of forecats in general and probability forecasts in particular (see, for example, Brier & Allen, 1951; Johnson, 1957; Murphy & Epstein, 1967). Of course, the relationships, on an absolute or even on an interval scale, between measures of the attributes "accuracy" and "value" can be expected to differ from measure to measure as well as from situation to situation. However, in comparative evaluation, in which meteorologists are concerned with the *relative* "accuracy" and "value" of forecasts prepared by two (or more) forecasters or forecast procedures, the relationships, on an *ordinal* scale, between measures of these attributes are of interest. Furthermore, in view of the difficulties inherent in any attempt on the part of meteorologists to determine the absolute or relative "value" of their forecasts directly, the relationships between the relative "accuracy" and relative "value" of such forecasts are of particular interest.

The problem of concern can be described as follows: If two forecasters, A and B say, prepare forecasts for the same situations and if A's forecasts are more "accurate" than B's forecasts, then what if any relationships exist, on an ordinal scale, between the "value" of A's and B's forecasts and what if any assumptions must be satisfied for these relationships to hold?

The purpose of this paper is to briefly describe the *preliminary* results of an investigation of the ordinal relationships between particular measures of the "accuracy" and "value" of probability forecasts. These and other related

[1] Supported in part by the National Science Foundation (Atmospheric Science Section) under Grant GA-1707.

[2] Contribution No. 193 from the Department of Meteorology and Oceanography, University of Michigan.

[3] This paper represents a revised version of Murphy (1971*a*).

[4] Present affiliation: Advanced Study Program, National Center for Atmospheric Research, Boulder, Colorado. The National Center for Atmospheric Research is sponsored by the National Science Foundation.

results are discussed in greater detail by Murphy (1971*b*). Since we are concerned in this paper with the relationships between measures of "accuracy" and "value" *in general*, we shall not be concerned with the formulation or description of evaluation measures which may be measures of "value" in *particular* situations. For examples and detailed discussions of such measures, refer to Epstein (1969), Murphy (1966, 1969, 1970, 1971*b*, 1972), and Stael von Holstein (1970).

In Section 2 we identify a measure of "accuracy", the *validity* measure, and two measures of "value", the *utility* and *expected-utility* measures, which are appropriate when the meteorologist's knowledge of the "values" of concern is complete and incomplete, respectively. The relationships, on an ordinal scale, between these measures are described in Section 3 for individual forecasts and collections of forecasts in general N-state $(N>2)$ and special two-state $(N=2)$ situations. The two-state, i.e., "weather—no weather", situation receives special attention because this situation is of particular interest to meteorologists and because certain ordinal relationships between these measures of "accuracy" and "value" hold in general in this situation. In Section 4 we present sample forecasts and consider both general and special situations which indicate the nature of the relationships and scope of the results. Section 5 consists of a brief summary and conclusion.

2. Measures of "Accuracy" and "Value"

We shall assume that the forecasts of concern are subjective probability forecasts defined on a set of N mutually exclusive and collectively exhaustive states $\{s_1, \ldots, s_N\}$. Then, a forecast is a vector $r=(r_1, \ldots, r_N)$, where $r_n \geqslant 0$ and $\Sigma_n r_n = 1$ $(n=1, \ldots, N)$, in which r_n expresses, in probabilistic terms, the meteorologist's "degree of belief", or judgment, that state s_n will obtain. (Alternatively, the forecasts of concern could have been formulated by two or more "objective" probability forecasting procedures.)

The set R of all possible forecasts, where

$$R=\left\{(r_1, \ldots, r_N) \,\middle|\, r_n \geqslant 0, \sum_n r_n = 1; n=1, \ldots, N\right\} \tag{1}$$

represents a regular (N-1)-dimensional simplex, in which each point corresponds to a particular forecast (and vice versa). A regular simplex is a unit line segment in two-state situations, an equilateral triangle in three-state situations, and a regular tetrahedron in four-state situations (see Pontryagin, 1952, pp. 10–12; see also Murphy, 1971*b*). Within the framework of the regular simplex R, the (euclidean) distance between two points r and r' is $D(r, r')$, where

$$D(r,r') = \left[\sum_{n=1}^{N} (r_n - r'_n)^2\right]^{\frac{1}{2}} \tag{2}$$

Note that if we assume that the vector $d=(d_1, \ldots, d_N)$, where d_n equals one if state s_n obtains and zero otherwise, denotes an observation, then the N vertices of the simplex R represent the N possible observations (as well as the N possible categorical forecasts).

2.1. *A measure of "accuracy"*

"Accuracy" is a general term which refers to the "correspondence" between forecasts and the relevant observations. When the forecasts of concern are *probability* forecasts, the term which describes the correspondence between forecasts and observations is *validity* (see, for example, Murphy & Epstein, 1967; Winkler & Murphy, 1968; Murphy & Winkler, 1970). At least two aspects of validity can be identified: (1) *primary* validity which refers to the correspondence between forecasts and observations on an individual basis and (2) *secondary* validity which refers to the correspondence between forecasts and observations on a collective basis (see Murphy & Winkler, 1970). In this paper we are concerned only with primary validity, which we shall refer to simply as validity.

As a measure of validity, we consider only the *validity measure* $V(r,d)$ (Winkler & Murphy, 1968; Murphy, 1971*b*), where

$$V(r,d) = 1-(1/2)\sum_{n=1}^{N} (r_n - d_n)^2 \tag{3}$$

Note that the range of $V(r,d)$ is the closed unit interval [0, 1] and that, if state s_j obtains (i.e., if $d_j=1$), $V(r,d)$ equals one when $r_j=1$ (r completely valid) and zero when $r_i=1$ $(i \neq j)$ (r "completely" invalid).

The measure $V(r,d)$ is *equivalent*, i.e., linearly related, to the probability score PS(r,d) (Brier, 1950) and the quadratic scoring rule $Q(r,d)$ (see Winkler & Murphy, 1968). In particular, $V(r,d) = 1 - (1/2)$ PS(r, d) and $V(r, d) = (1/2)\,[1 + Q(r,d)]$. Thus, the measure $V(r, d)$ is a *strictly proper* scoring rule (Winkler & Murphy, 1968).

Within the geometrical framework of the simplex R, the distance $D(r, d)$ between r and d is a natural measure of validity. Note, from (2) and (3), that

$$V(r, d) = 1 - (1/2)\, D^2(r, d) \tag{4}$$

Thus, the measure $V(r, d)$ is also equivalent to the square of the distance measure $D(r, d)$.

2.2. *Measures of "value"*

As Fishburn (1964, p. 2) indicates, "value ... has meaning only ... within the context of a decision situation". The decision situations of concern in this paper can be described in terms of their primitive elements: a decision maker (DM); a set of M admissible actions $\{a_1, \ldots, a_M\}$; a set of N mutually exclusive and collectively exhaustive states $\{s_1, \ldots, s_N\}$; a set of MN consequences $\{o_{11}, \ldots, o_{MN}\}$ (consequence o_{mn} results when the DM selects action a_m and state s_n obtains); a matrix $U = (u_{mn})$ of the DM's "values", or utilities, on the consequences (the utilities express the DM's preferences for the consequences); a vector $p = (p_1, \ldots, p_N)$, where $p_n \geq 0$ and $\Sigma_n p_n = 1$ $(n = 1, \ldots, N)$, which expresses, in probabilistic terms, the DM's judgments on the states; and a decision rule, according to which the DM selects the action a_m for which his expected utility u_m, where

$$u_m = \sum_{n=1}^{N} p_n u_{mn} \tag{5}$$

is a maximum. Note that we assume that the DM's set of states is identical to the meteorologist's set of states. We also assume that the DM adopts the meteorologist's forecast r as his judgment p, in which case u_m in (5) becomes

$$u_m = \sum_{n=1}^{N} r_n u_{mn} \tag{6}$$

These decision situations can be depicted within the framework of the simplex R. Specifically, the set R is subdivided into M convex subsets R_m, where

$$R_m = \Big\{(r_1, \ldots, r_N) \,\Big|\, r_n \geq 0, \sum_n r_n = 1, u_m > u_{m'}$$
$$\text{for all } m' \neq m;\ m, m' = 1, \ldots, M; n = 1, \ldots, N\Big\} \tag{7}$$

The subset R_m represents the set of all forecasts for which the DM prefers action a_m to the other M-1 actions. [Note that the set of forecasts for which $u_m = u_{m'}$, i.e., for which the DM is *indifferent* between actions a_m and $a_{m'}$, is not contained in subsets R_m or $R_{m'}$. Such sets, which represent (N-2)-dimensional indifference hyperplanes, are not of particular importance and are not considered in this paper.] Hereafter, $r \in R_m$ will denote that the forecast r "belongs to" the subset R_m.

Note that, within this framework for decision making, in which "value" is expressed in terms of utility, the consequence that results, and hence the utility of that consequence to the DM, depends upon the forecast r. Thus, the utility of the consequence of concern is a natural measure of the "value" of r to the DM. If we denote this utility measure by $U(r, d)$, then

$$U(r, d) = \sum_{m=1}^{M} \delta_m \left(\sum_{n=1}^{N} d_n u_{mn} \right) \tag{8}$$

where

$$\delta_m = \begin{matrix} 1 & \text{if } r \in R_m \\ 0 & \text{otherwise} \end{matrix} \tag{9}$$

[When $r \in R_m$ and $d_n = 1$, $U(r, d)$ is simply u_{mn}.] Now, if we assume that the most and least preferred consequences can always be identified, then the utility matrix $U = (u_{mn})$ can be transformed into a *standard* utility matrix $U' = (u'_{mn})$, in which $0 \leq u'_{mn} \leq 1$ (see Murphy, 1972). Hereafter, we shall assume that the matrix $U = (u_{mn})$ is a standard utility matrix, in which case the range of the utility measure $U(r, d)$ is the closed unit interval [0,1]. The measure $U(r, d)$ is a *proper*, but not a strictly proper, scoring rule (see, for example, Murphy, 1971*b*; Stael von Holstein, 1970).

Since the measure of "value", $U(r, d)$, involves the DM's utilities, the meteorologist's knowledge of these utilities must be considered. We identify two possible states of the meteorologist's knowledge: (1) *State I*—the meteorologist's knowledge is incomplete and (2) *State II*—the meteorologist's knowledge is complete. The specification of the appropriate measures of "value" for these states of knowledge can be described as follows: (1) *State I*. The meteorologist's knowledge of the utilities is assumed to be expressed in terms of a probability distribution $F(U) = F(u_{11}, \ldots, u_{MN})$. (The subjective interpretation of probability, the interpretation to which we subscribe, prescribes that the meteorologist can *always* express his knowledge of the DM's utilities in probabilistic terms.) A natural measure of the "value" of the forecast r to the DM in this situation is the expected-utility measure $EU(r, d)$, where

$$EU(r,d) = \int_R \left[\sum_{m=1}^{M} \delta_m \left(\sum_{n=1}^{N} d_n u_{mn} \right) \right] dF(U) \qquad (10)$$

The range of the measure $EU(r, d)$ is, of course, the closed unit interval [0,1]. With regard to the "form" of the distribution $F(U)$, we identify two substates of this state of the meteorologist's knowledge: (i) *substate* (*a*)—$F(U)$ "defines" a discrete probability function or a probability density function which is not strictly positive and (ii) *substate* (*b*)—$F(U)$ defines a strictly positive probability density function. The measure $EU(r, d)$ is a *proper* and *strictly proper* scoring rule in substates (*a*) and (*b*), respectively (see Murphy, 1971*b*). (2) *State II*. The meteorologist's knowledge of the utilities is assumed to be complete, i.e., the meteorologist is assumed to be able to specify the utilities. Thus, the State II situation is a special case of the State I situation, in which the probability distribution $F(U)$ is concentrated at the "point" $U = (u_{mn})$. In such a situation the expected-utility measure $EU(r, d)$ "reduces to" the utility measure $U(r, d)$, and the latter represents the appropriate measure of "value".

Heretofore, we have assumed, at least implicitly, that the meteorologist's forecast is utilized by only one DM and that, in this "one-on-one" situation, the distribution $F(U)$ expresses the meteorologist's knowledge of the DM's utilities. In "one-on-many" situations in which the meteorologist's forecast is utilized by many DMs, the distribution $F(U)$ (in State I) can be considered to represent the distribution of the different DMs' utilities (see Murphy, 1966; Murphy & Winkler, 1970).

3. Ordinal Relationships: Results

We shall denote the distance between A's forecasts and the relevant observations, as measured by $D(r, d)$, by DA; the validity of A's forecasts, as measured by $V(r, d)$, by VA; and the utility and expected utility of A's forecasts, as measured by $U(r, d)$ and $EU(r, d)$, by UA and EUA, respectively. Similarly, we shall denote the distance, validity, and utility and expected utility of B's forecasts by DB, VB, UB and EUB, respectively.

3.1. *Individual forecasts*

The validity measure $V(r, d)$ is, from (4), a *strictly decreasing* function of the distance $D(r, d)$ between r and d (see Section 4 and Figs. 1 and 4 for the two-state and three-state situations, respectively). That is, if $DA > DB$, then $VA < VB$. (We shall consider only the situation in which $DA > DB$. If $DA < DB$, then we can relabel the forecasters. Note that forecasts for which $DA = DB$ are not considered.)

The utility measure $U(r, d)$, a step function in N-1 dimensions, is *not*, in general, a decreasing (or an increasing) function of the distance $D(r, d)$. [This statement and other similar statements in this section have been proved by Murphy (1971*b*).] Specifically, if $DA > DB$, then either $UA \geqslant UB$ or $UA \leqslant UB$. However, the measure $U(r, d)$ is a *decreasing* function of *directed* distance (see Section 4 and Figs. 3 and 6 for examples in the two-state and three-state situations, respectively). The directed distance is the (euclidean) distance $D(r, d)$ between r and d on the line segment through r and the vertex of the simplex R which represents d. We shall refer to two forecasts as *comparable* if they correspond to points on the same line segment through the vertex which represents d. [When two forecasts are comparable certain relationships exist among the ratios of the components of the forecasts. For example, in a three-state situation in which state s_j obtains, i.e., in which $d_j = 1$, the forecasts r^A and $r^B (r^A \neq r^B)$ are comparable if $r_i^A/r_i^B = r_k^A/r_k^B = (1 - r_j^A)/(1 - r_j^B)$ ($i, j, k = 1, 2, 3$; $i \neq k$; $i, k \neq j$).] Thus, if A's and B's forecasts are comparable and if $DA > DB$, then $UA \leqslant UB$.

The expected-utility measure $EU(r, d)$, in substate I(*a*), is also a *decreasing* function of directed distance. However, in substate I(*b*), i.e., when the distribution $F(U)$ is such that a strictly positive density function exists, the measure $EU(r, d)$ is a *strictly decreasing* function of directed distance. That is, if A's and B's forecasts are comparable and if $DA > DB$, then $EUA < EUB$.

The relationships, on an ordinal scale, between the measures of concern are summarized in Result 1.

Result 1 (N States, One Forecast)

Assumption. A's and B's forecasts are comparable.

Results. (1) *State I:* (i) substate (a)—If $V\text{A} > V\text{B}$, then $EU\text{A} \geqslant EU\text{B}$. (ii) substate (b)—If $V\text{A} > V\text{B}$, then $EU\text{A} > EU\text{B}$. (2) *State II:* If $V\text{A} > V\text{B}$, then $U\text{A} \geqslant U\text{B}$.

Thus, for individual, comparable forecasts, the validity measure $V(r, d)$ represents an ordinal measure on utility and expected utility, i.e., the measure $V(r, d)$ is able to order, or rank, forecasters according to the utility or expected utility of their forecasts.

Unfortunately, A's and B's forecasts are not often comparable. However, on many occasions the forecasts may be "nearly" comparable, i.e., the forecasts may correspond to points on two line segments which are "close" to one another. On such occasions the measure $V(r, d)$ represents an ordinal measure on utility and expected utility for most, but not all, decision situations and states of the meteorologist's knowledge (or distributions of the DMs' utilities).

Note that in *two-state* $(N = 2)$ situations, in which the utility measure $U(r, d)$ is a step function in one dimension with M-1 steps, the simplex R is a (unit) line segment and, as a result, distance and directed distance are identical. Thus, *all forecasts are comparable in two-state situations.* Therefore, the relationships under Result 1 hold, in general, in two-state situations. (Since strictly proper scoring rules are strictly decreasing or increasing functions of distance in two-state situations, this result holds for all such scoring rules. If the scoring rule of concern is a strictly increasing function of distance, then certain inequalities in Result 1 must be reversed.)

3.2. *Collections of forecasts*

For collections of $K (K > 1)$ forecasts, we shall be concerned with the relationships, on an ordinal scale, between the *average* "accuracy" and *average* "value" of the forecasts.

Although the validity $V(r, d)$ represents an ordinal measure on the utility $U(r, d)$ and expected utility $EU(r, d)$ for *individual*, comparable forecasts, the average validity $\bar{V}(r, d)$ is *not*, in general, an ordinal measure on the average utility $\bar{U}(r, d)$ or average expected utility $\overline{EU}(r, d)$ for *collections* of individually comparable forecasts. However, we can identify two special situations, in which, under certain assumptions, average validity is an ordinal measure on average utility and average expected utility: (1) *Situation 1*—K forecasts for K (different) decision situations and (2) *Situation 2*—K forecasts for one decision situation (i.e., K forecasts on K different occasions for the same decision situation).

The relationships, on an ordinal scale, between the measures of concern in Situations 1 and 2 are summarized in Results 2.1 and 2.2, respectively.

Result 2.1 (N States, K Forecasts, K Decision Situations)

Assumptions. (1) $V\text{A}_k > V\text{B}_k$ for all k ($k = 1, \ldots, K$) and (2) A's and B's forecasts are individually comparable.

Results. The relationships under Result 1 hold for average validity, average utility, and average expected utility.

Result 2.2 (N States, K Forecasts, One Decision Situation)

Assumptions. (1) $V\text{A}_{(k)} > V\text{B}_{(k)}$ for all k ($k = 1, \ldots, K$), where $V\text{A}_{(k)}$ and $V\text{B}_{(k)}$ represent the *order statistics* (Wilks, 1962, pp. 234–245) of the validities of A's and B's forecasts, respectively, and (2) A's and B's forecasts are individually comparable. (Note that in this situation the forecasts are assumed to be comparable *after* the order statistics have been obtained.)

Results. The relationships under Result 1 hold for average validity, average utility, and average expected utility.

Thus, for collections of individually comparable forecasts, the average validity $\bar{V}(r, d)$ represents an ordinal measure on the average utility $\bar{U}(r, d)$ and average expected utility $\overline{EU}(r, d)$ in these two *special* situations.

The assumption that $V\text{A}_k > V\text{B}_k$ for all k, which yields Result 2.1, is, of course, a very restrictive assumption. That is, collections of A's and B's forecasts will not often satisfy this assumption. The assumption that $V\text{A}_{(k)} > V\text{B}_{(k)}$ for all k, which yields Result 2.2, is somewhat less restrictive. However, in this special situation we also assume that the forecasters are concerned with the same decision situation on the K occasions (and that the forecasts are comparable after the order statistics have been obtained).

Recall that all forecasts are comparable in

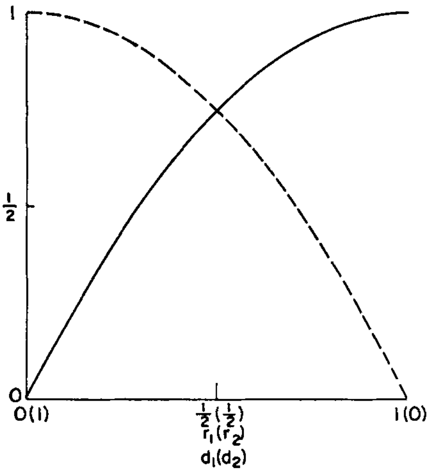

Fig. 1. The validity measure $V(r, d)$ in the two-state situation. The solid (dashed) curve denotes $V(r, d)$ when state $s_1(s_2)$ obtains [i.e., when $d_1(d_2) = 1$].

<u>Decision Situation</u>

Utility Matrix $\underline{U}$

		States		Expected Utility
		s_1	s_2	u
	a_1	1	0	$u_1 = r_1$
Actions	a_2	2/3	1/3	$u_2 = (2/3)r_1 + (1/3)r_2$
	a_3	1/3	2/3	$u_3 = (1/3)r_1 + (2/3)r_2$
Probability Vector	$\underline{r}$	r_1	r_2	

<u>Expected-Utility Decision Rule</u>

a_1 if $\underline{r} \in R_1$ (i.e., if $r_1 > 2/3$)

a_2 if $\underline{r} \in R_2$ (i.e., if $1/2 < r_1 < 2/3$)

a_3 if $\underline{r} \in R_3$ (i.e., if $r_1 < 1/2$)

<u>Simplex Representation</u>

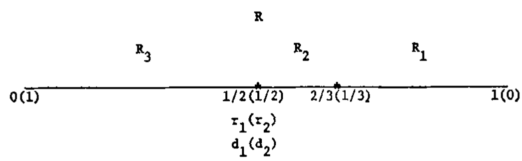

Fig. 2. A particular three-action–two-state decision situation, the expected-utility decision rule in this situation, and the simplex representation of this situation.

two-state situations. Thus, the relationships under Results 2.1 and 2.2 hold under somewhat less restrictive assumptions in such situations.

4. Ordinal Relationships: Illustrations

In this section we present sample forecasts and consider general and special situations which illustrate the nature of the relationships and scope of the results described in Section 3. For the purposes of this presentation, we consider two-state ($N = 2$) and N-state ($N > 2$) situations separately in 4.1 and 4.2, respectively.

4.1. *Two-state ($N = 2$) situations*

Recall that in two-state situations all forecasts are comparable (see 3.1).

In Fig. 1 we depict the validity measure $V(r, d)$ in the two-state situation. [The coordinate system used to depict forecasts in Fig. 1 is a "length" coordinate system, in which the distances between the point which represents a forecast and the end points of the unit line segment are the two components of the forecast. In N-state situations the length coordinate system becomes a "content" coordinate system, in which the distances between the point which represents a forecast and the "faces" of the $(N-1)$-dimensional simplex are the N components of the forecast (see Springer, 1964, pp. 119–122).] Note that, as indicated in 3.1, the measure $V(r, d)$ is a strictly decreasing function of the distance $D(r, d)$. That is, $V(r, d)$ decreases as the distance between r and d increases.

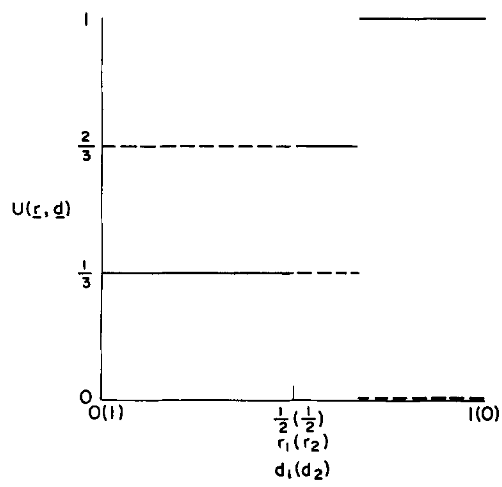

Fig. 3. The utility measure $U(r, d)$ in the decision situation depicted in Fig. 2. The solid (dashed) lines denote $U(r, d)$ when state $s_1(s_2)$ obtains [i.e., when $d_1(d_2) = 1$].

Table 1. *Sample collections of ten forecasts prepared by forecasters A and B, the relevant observations, and the distances, validities, and utilities of the forecasts according to measures $D(r, d)$, $V(r, d)$, and $U(r, d)$, respectively, for two-state situations (the untilies relate to the decision situation depicted in Fig. 2)*

Forecast no. k	Forecasts $r=(r_1, r_2)$ r_k^A	r_k^B	Observation $d=(d_1, d_2)$ d_k	Distance measure $D(r, d)$ DA_k	DB_k	Validity measure $V(r, d)$ VA_k	VB_k	Utility measure $U(r, d)$ UA_k	UB_k
1	(0.20, 0.80)	(0.25, 0.75)	(0,1)	0.20	0.25	0.98000	0.96875	2/3	2/3
2	(0.35, 0.65)	(0.30, 0.70)	(0,1)	0.35	0.30	0.93875	0.95500	2/3	2/3
3	(0.80, 0.20)	(0.60, 0.40)	(1,0)	0.20	0.40	0.98000	0.92000	1	2/3
4	(0.60, 0.40)	(0.65, 0.35)	(0,1)	0.60	0.65	0.82000	0.78875	1/3	1/3
5	(0.90, 0.10)	(0.85, 0.15)	(1,0)	0.10	0.15	0.99500	0.98875	1	1
6	(0.25, 0.75)	(0.20, 0.80)	(0,1)	0.25	0.20	0.96875	0.98000	2/3	2/3
7	(0.05, 0.95)	(0.10, 0.90)	(0,1)	0.05	0.10	0.99875	0.99500	2/3	2/3
8	(0.15, 0.85)	(0.20, 0.80)	(0,1)	0.15	0.20	0.98875	0.98000	2/3	2/3
9	(0.85, 0.15)	(0.75, 0.25)	(1,0)	0.15	0.25	0.98875	0.96875	1	1
10	(0.40, 0.60)	(0.45, 0.55)	(0,1)	0.40	0.45	0.92000	0.89875	2/3	2/3
Average				0.245	0.295	0.95788	0.94438	0.733	0.700

In Fig. 2 we depict a particular three-action–two-state decision situation. The utility measure $U(r, d)$ in this decision situation is depicted in Fig. 3. Note that, as indicated in 3.1, the measure $U(r, d)$, a step function with two steps, is a decreasing function of the distance $D(r, d)$. That is, $U(r, d)$ is constant or decreases as the distance between r and d increases.

In Table 1 we present sample collections of forecasts prepared by forecasters A and B for two-state situations. Consider A's and B's forecasts on the first occasion ($k=1$). Note that $VA_1 > VB_1$. Thus, $UA_1 \geqslant UB_1$ for all decision situations. In particular, for the decision situation depicted in Fig. 2, $UA_1 = UB_1$ $(=2/3)$. Now, consider A's and B's forecasts on the third occasion ($k=3$). Since $VA_3 > VB_3$, $UA_3 \geqslant UB_3$ for all decision situations. For the decision situation depicted in Fig. 2, UA_3 $(=1) > UB_3$ $(=2/3)$. Thus, as Result 1 indicates, the validity measure $V(r, d)$ represents an ordinal measure on utility for *individual* (comparable) forecasts. A similar relationship holds for expected utility (in which case the utilities in the decision situation of concern are assumed to be random variables with a joint probability distribution).

Table 2. *The validities and utilities of the forecasts presented in Table 1 depicted in terms of the order statistics of the validities*

Order statistic (k)	A's forecasts: Forecast no. k	Validity $VA_{(k)}$	Utility $UA_{(k)}$	B's forecasts: Forecast no. k	Validity $VB_{(k)}$	Utility $UB_{(k)}$
1	7	0.99875	2/3	7	0.99500	2/3
2	5	0.99500	1	5	0.98875	1
3	8	0.98875	2/3	6	0.98000	2/3
4	9	0.98875	1	8	0.98000	2/3
5	1	0.98000	2/3	1	0.96875	2/3
6	3	0.98000	1	9	0.96875	1
7	6	0.96875	2/3	2	0.95500	2/3
8	2	0.93875	2/3	3	0.92000	2/3
9	10	0.92000	2/3	10	0.89875	2/3
10	4	0.82000	1/3	4	0.78875	1/3
Average		0.95788	0.733		0.94438	0.700

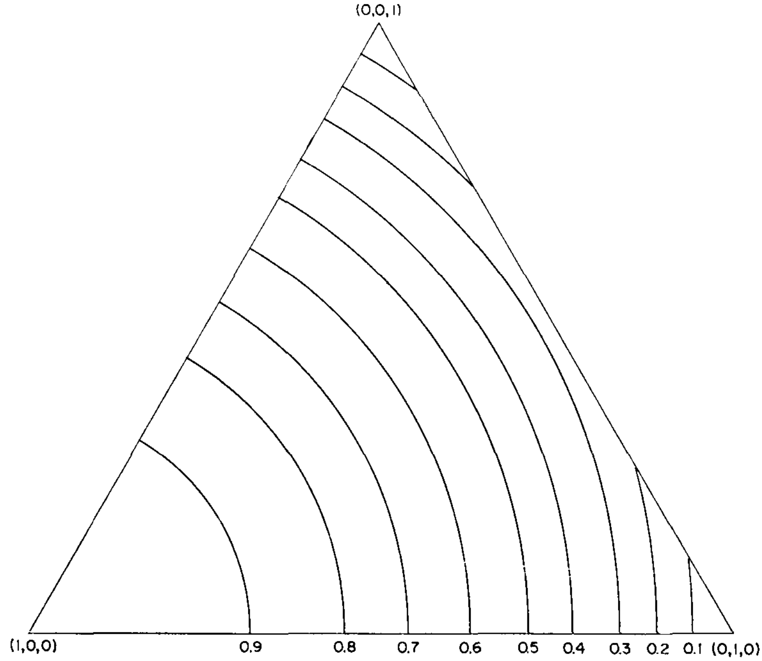

Fig. 4. The validity measure $V(r, d)$ in the three-state situation when state s_1 obtains (i.e., when $d_1 = 1$).

Now, consider the forecasts presented in Table 1 as two *collections* of forecasts. Note that $VA_k \ngtr VB_k$ for all k ($k = 1, \ldots, 10$). Thus, if we are concerned with Situation 1, i.e., ten forecasts for ten different decision situations (see 3.2), then either $\bar{U}\mathrm{A} \geq \bar{U}\mathrm{B}$ or $\bar{U}\mathrm{A} \leq \bar{U}\mathrm{B}$ (even though $\bar{V}\mathrm{A} > \bar{V}\mathrm{B}$). In Table 2 we present the validities and utilities of A's and B's forecasts in terms of the order statistics of the validities. Note that $VA_{(k)} > VB_{(k)}$ for all k ($k = 1, \ldots, 10$). Thus, if we are concerned with Situation 2, i.e., ten forecasts for the same decision situation (see 3.2), then $\bar{U}\mathrm{A} \geq \bar{U}\mathrm{B}$. In particular, for the decision situation depicted in Fig. 2, $\bar{U}\mathrm{A}$ ($= 0.733$) $> \bar{U}\mathrm{B}$($= 0.700$). Thus, as Results 2.1 and 2.2 indicate, the average validity $\bar{V}(r, d)$ represents an ordinal measure on the average utility $\bar{U}(r, d)$ [and average expected utility $\overline{EU}(r, d)$] for *collections* of (individually comparable) forecasts in these two special situations. However, as indicated in 3.2, collections of forecasts, even in two-state situations, do not often satisfy the assumptions which yield Results 2.1 and 2.2.

4.2. *N-state* ($N > 2$) *situations*

Recall that forecasts are comparable in N-state situations only if the forecasts correspond to points on the same line segment through the vertex of the simplex R which represents the observation d (see 3.1). We shall consider only three-state ($N = 3$) situations.

In Fig. 4 we depict the validity measure $V(r, d)$ in the three-state situation when state s_1 obtains (i.e., when $d_1 = 1$). (The isopleths of validity are, of course, arcs of circles whose center is the vertex for which $d_1 = 1$.) The

Decision Situation

Utility Matrix $\underline{U}$

		States			Expected Utility
		s_1	s_2	s_3	u
Actions	a_1	1	0	1/3	$u_1 = r_1 + (1/3)r_3$
	a_2	2/3	5/6	1/3	$u_2 = (2/3)r_1 + (5/6)r_2 + (1/3)r_3$
	a_3	1/3	2/3	1/2	$u_3 = (1/3)r_1 + (2/3)r_2 + (1/2)r_3$
	a_4	1/6	1/3	2/3	$u_4 = (1/6)r_1 + (1/3)r_2 + (2/3)r_3$
Probability Vector	$\underline{r}$	r_1	r_2	r_3	

Expected-Utility Decision Rule

a_m if $\underline{r} \in R_m$ $(m=1,2,3,4)$

Fig. 5. (*a*) A particular four-action–three-state decision situation and the expected-utility decision rule in this situation.

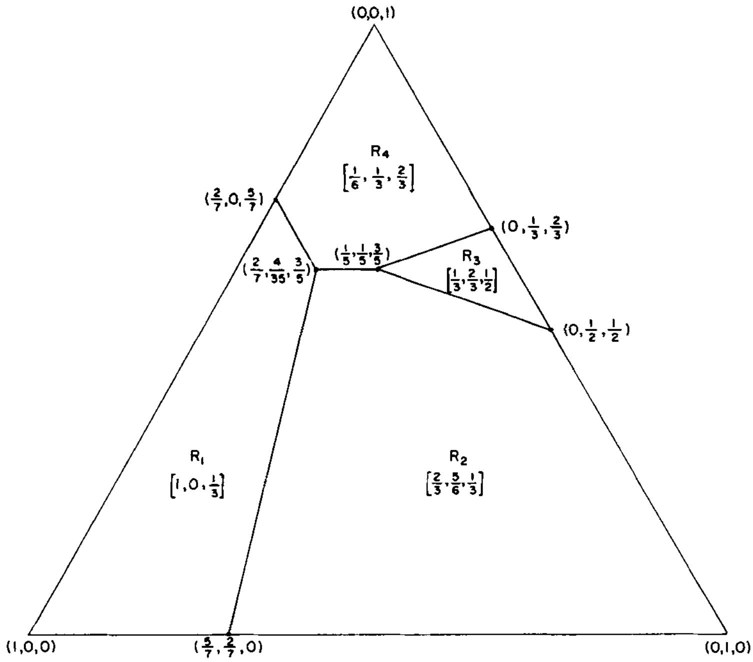

Fig. 5. (*b*) The simplex representation of the decision situation depicted in Fig. 5*a* and the utility measure $U(r, d)$ in this situation.

measure $V(r, d)$, as indicated in 3.1, is a strictly decreasing function of the distance $D(r, d)$. Since the measure $V(r, d)$ is a symmetric function, $V(r, d)$ when $d_2(d_3) = 1$ can be determined from Fig. 4 simply by interchanging the labels on the vertices corresponding to states s_1 and $s_2(s_3)$.

In Fig. 5*a* we depict a particular four-action–three-state decision situation. The simplex representation of this decision situation is depicted in Fig. 5*b*. The utility measure $U(r, d)$ in this situation, a "plateau" function on the set R, is "depicted" numerically in Fig. 5*b*. If, for example, state s_1 obtains (i.e., $d_1 = 1$), then the "height" of the plateau is 1, 2/3, 1/3, and 1/6 in subsets R_1, R_2, R_3, and R_4, respectively. Note that the measure $U(r, d)$ is a decreasing function of directed distance, i.e., $U(r, d)$ is constant or decreases as the distance $D(r, d)$ between r and d on any line segment through d increases (see 3.1).

In Table 3 we present sample forecasts pre-

Table 3. *Sample forecasts prepared by forecasters A and B, the relevant observations, and the distances, validities, and utilities of the forecasts according to measures $D(r, d)$, $V(r, d)$, and $U(r, d)$, respectively, for three-state situations (the utilities relate to the decision situation depicted in Fig. 5)*

Forecast no.	Forecasts $r = (r_1, r_2, r_3)$		Observation $d = (d_1, d_2, d_3)$	Distance measure $D(r, d)$		Validity measure $V(r, d)$		Utility measure $U(r, d)$	
k	r_k^A	r_k^B	d_k	DA_k	DB_k	VA_k	VB_k	UA_k	UB_k
1	(0.05, 0.40, 0.55)	(0.30, 0.15, 0.55)	(0, 1, 0)	0.8155	1.0559	0.6675	0.4425	5/6	5/6
2	(0.15, 0.80, 0.05)	(0.30, 0.60, 0.10)	(0, 1, 0)	0.2550	0.5099	0.9675	0.8700	5/6	5/6
3	(0.15, 0.70, 0.15)	(0.20, 0.50, 0.30)	(0, 1, 0)	0.3674	0.6164	0.9325	0.8100	2/3	5/6
Average				0.4793	0.7274	0.8558	0.7075	0.778	0.833

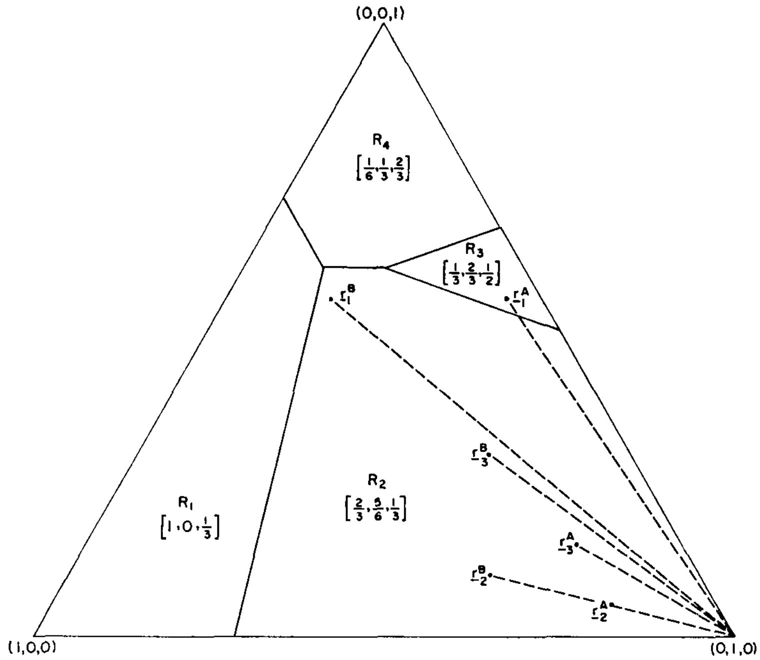

Fig. 6. The representation of the forecasts presented in Table 3 within the framework of the simplex R and the decision situation depicted in Fig. 5.

pared by forecasters A and B for three-state situations. Consider A's and B's forecasts on the first occasion ($k = 1$). Note, in Fig. 6, that the points r_1^A and r_1^B do *not* lie on the same line segment through d ($d_2 = 1$). Thus, A's and B's forecasts on this occasion are not comparable. Therefore, in general, either $UA_1 \geqslant UB_1$ or $UA_1 \leqslant UB_1$ (even though $VA_1 > VB_1$). In particular, for the decision situation depicted in Fig. 5, UA_1 ($=2/3$) $< UB_1$($=5/6$) even though VA_1 ($=0.6675$) $> VB_1$ ($=0.4425$). Now, consider A's and B's forecasts on the second occasion ($k = 2$). Note, in Fig. 6, that the points r_2^A and r_2^B lie on the same line segment through d ($d_2 = 1$). Thus, A's and B's forecasts on this occasion are comparable, and, since $VA_2 > VB_2$, $UA_2 \geqslant UB_2$ for all decision situations. For the decision situation depicted in Fig. 5, $UA_2 = UB_2$ ($=5/6$). Thus, as indicated in 3.1, the validity measure $V(r, d)$ represents an ordinal measure on utility for *individual*, comparable forecasts. A similar relationship holds for expected utility. Unfortunately, as indicated in 3.1, forecasts are not often comparable in N-state situations. However, if the forecasts of concern are nearly comparable, then these ordinal relationships can be expected to hold in most situations. For example, A's and B's forecasts on the third occasion ($k = 3$) correspond to points on two line segments through d which are close to one another. Thus, since $VA_3 > VB_3$, $UA_3 \geqslant UB_3$ for most, *but not all*, decision situations.

With regard to *collections* of forecasts, the forecasts presented in Table 3 indicate that, unless the forecasts of concern are individually comparable, average validity is not an ordinal measure on average utility (and average expected utility). Note that the forecasts on the first and third occasions are not comparable. Thus, even though $VA_k > VB_k$ for all k($k = 1$, 2, 3) (and, as a result, $\overline{VA} > \overline{VB}$), in general, either $\overline{UA} \geqslant \overline{UB}$ or $\overline{UA} \leqslant \overline{UB}$. In particular, for the decision situation depicted in Fig. 5, $\overline{UA}$($=0.778$) $< \overline{UB}$($=0.833$).

5. Summary and Conclusion

We have described the preliminary results of an investigation of the relationships, on an ordinal scale, between particular measures of the "accuracy" (the *validity* measure) and "value" (the *utility* and *expected-utility* measures) of probability forecasts. The results indicate that the validity measure represents an ordinal measure on utility and expected utility for *individual*, comparable forecasts. That is, the validity measure is able to order, or rank, forecasters or forecast procedures according to the utility or expected utility of their forecasts for individual, comparable forecasts. Although forecasts in N-state ($N > 2$) situations are not often comparable, this result can be expected to hold in most situations in which the forecasts are nearly comparable. In addition, all forecasts are comparable in two-state ($N = 2$) situations. For *collections* of comparable forecasts the average validity is an ordinal measure on average utility and average expected utility in special situations, in which certain rather restrictive assumptions concerning the relative validities of the individual forecasts in the collections are satisfied.

We are presently investigating (1) whether such results can be obtained in these and other situations under less restrictive assumptions, (2) the extent to which the relationships, on an ordinal scale, between other measures of "accuracy" and "value" differ from those described herein, and (3) the nature of the relationships between such measures on scales other than an ordinal scale.

Acknowledgements

The author would like to acknowledge the comments of Edward S. Epstein and Robert L. Winkler on earlier versions of this paper.

REFERENCES

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

Brier, G. W. & Allen, R. A. 1951. Verification of weather forecasts. *Compendium of Meteorology*, pp. 841–848. American Meteorological Society, Boston.

Epstein, E. S. 1969. A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor. 8*, 985–987.

Fishburn, P. C. 1964. *Decision and value theory.* Wiley, New York. 451 pp.

Johnson, D. H. 1957. Forecast verification: a critical survey of the literature. Air Ministry, London. Meteorological Research Committee, M.R.P. No. 1056. 40 pp.

Murphy, A. H. 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.

Murphy, A. H. 1969. Measures of the utility of probabilistic predictions in cost–loss ratio decision situations in which knowledge of the cost–loss ratios is incomplete. *J. Appl. Meteor. 8*, 863–873.

Murphy, A. H. 1970. The ranked probability score and the probability score: a comparison. *Mon. Wea. Rev. 98*, 917–924.

Murphy, A. H. 1971*a*. On the relationship between the "accuracy" and the "value" of probability forecasts. *Preprints of Papers Presented at the International Symposium on Probability and Statistics in the Atmospheric Sciences*, pp. 94–99. American Meteorological Society, Boston.

Murphy, A. H. 1971*b*. The evaluation of probabilistic predictions in meteorology. Ann Arbor, University of Michigan, Department of Meteorology and Oceanography, manuscript, unpublished.

Murphy, A. H. 1972. Measures of the value of probability forecasts in meteorological decision situations. Boulder, Colo., National Center for Atmospheric Research, Advanced Study Program, manuscript, unpublished.

Murphy, A. H. & Epstein, E. S. 1967. Verification of probabilistic predictions: a brief review. *J. Appl. Meteor. 6*, 748–755.

Murphy, A. H. & Winkler, R. L. 1970. Scoring rules in probability assessment and evaluation. *Acta Psychol. 34*, 273–286.

Pontryagin, L. S. 1952. *Foundations of combinatorial topology.* Graylock Press, Rochester, N.Y. 99 pp.

Springer, C. E. 1964. *Geometry and analysis of projective spaces.* W. H. Freeman, San Francisco. 299 pp.

Stael von Holstein, C.-A. S. 1970. A family of strictly proper scoring rules which are sensitive to distance. *J. Appl. Meteor. 9*, 360–364.

Wilks, S. S. 1962. *Mathematical statistics.* Wiley, New York. 644 pp.

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

ПОРЯДКОВЫЕ СООТНОШЕНИЯ МЕЖДУ МЕРАМИ «ТОЧНОСТИ» И «ЦЕННОСТИ» ВЕРОЯТНОСТНЫХ ПРОГНОЗОВ. ПРЕДВАРИТЕЛЬНЫЕ РЕЗУЛЬТАТЫ 1, 2, 3

Описываются предварительные результаты исследования соотношений в порядковом масштабе между мерами «точности» (мера оправдываемости) и «ценности» (мера полезности и ожидаемой полезности) вероятностных прогнозов. Для индивидуальных сравнимых прогнозов мера оправдываемости представляет собою порядковую меру полез-

ности и ожидаемой полезности, т. е. мера оправдываемости способна упорядочить или расставить по рангу прогнозистов и (или) прогностические процедуры в соответствии с полезностью или ожидаемой полезностью прогнозов. Хотя прогнозы в общем случае не часто сравнимы, все прогнозы сравнимы по двум состояниям, т. е. по ситуациям «погода–непогода». К тому же можно ожидать, что эти соотношения выполняются в большинстве ситуаций, в которых прогнозы почти сравнимы. Для совокупности индивидуально сравнимых прогнозов подобные соотношения имеют место между средней оправдываемостью и средней полезностью и ожидаемой полезностью в некоторых особых ситуациях. Природа соотношений и пределы применимости результатов как в общем случае, так и в специальных ситуациях устанавливается с помощью выборочных коллекций прогнозов.