"Good" Probability Assessors[1]

ROBERT L. WINKLER

Indiana University, Bloomington

AND ALLAN H. MURPHY[2]

University of Michigan, Ann Arbor

(Manuscript received 8 December 1967, in revised form 10 May 1968)

ABSTRACT

Since a meteorologist's predictions are subjective, a framework for the evaluation of meteorological probability assessors must be consistent with the theory of subjective probability. Such a framework is described in this paper. First, two standards of "goodness," one based upon normative considerations and one based upon substantive considerations, are proposed. Specific properties which a meteorologist's assessments should possess are identified for each standard. Then, several measures of "goodness," or scoring rules, which indicate the extent to which such assessments possess certain properties, are described. Finally, several important uses of these scoring rules are considered.

1. Introduction

Meteorologists today often express their predictions in probabilistic form. The precipitation probability statements in the Weather Bureau's public weather forecasts provide perhaps the best examples of such predictions. However, probabilistic predictions of a variety of meteorological elements have been prepared on an experimental basis [e.g., Enger *et al.*, 1962; Sanders, 1963; Federal Aviation Agency/U. S. Weather Bureau, 1965; refer to Murphy (1968a) for a comprehensive bibliography of related material]. Sound meteorological (scientific) and operational (economic) reasons exist, of course, for expressing meteorological predictions in probabilistic form (e.g., Malone, 1957; refer to Section 2).

We assume, in this paper, that the probabilistic predictions issued by meteorologists are subjective, i.e., that each prediction expresses the "judgment" of a particular meteorologist on a particular occasion.[3] The theory of subjective probability, as developed by de Finetti (1937) and Savage (1954), provides a framework within which an individual, e.g., a meteorologist, can quantify his judgments in this manner. Probabilities, in this framework, are interpreted in terms of degrees of belief.[4] Specifically, a subjective probability measures the confidence that an individual has in the truth of a particular proposition, e.g., a meteorologist's confidence in the statement "precipitation tomorrow." This confidence may, of course, vary from individual to individual. The subjective framework does not admit the existence of a "correct" (in the sense of universal) probability.

We are concerned, in this paper, with formulating a framework for evaluating, i.e., for determining the "goodness" of, the meteorologists who assess these probabilities, the "assessors."[5] Since the probabilities of concern are subjective, such a framework must, of course, be consistent with the theory of subjective probability. Two standards of "goodness," one based upon normative (probabilistic) considerations and one based upon substantive (meteorological and operational) considerations, and several properties for each standard, properties which a meteorologist's assessments[6] should possess, are identified in Section 2. Several measures of "goodness," or scoring rules, which indicate the extent to which the assessments possess certain properties, are described in Section 3. In Section 4 three important uses of such scoring rules are considered: 1) to encourage meteorologists to make

[1] Supported, in part, by the National Science Foundation (Atmospheric Sciences Section) under Grant GA-906.

[2] Present affiliation: Mathematical Sciences Department, The Travelers Research Center, Inc., Hartford, Conn.

[3] Such an assumption is, we believe, consistent with present meteorological practice (e.g., Hughes, 1965; C. F. Roberts, 1965). Note that this assumption places no particular restriction on the manner in which a meteorologist formulates his predictions. That is, the meteorologist may utilize objective as well as subjective procedures.

[4] A meteorologist's judgment, then, represents his *true* belief, while the prediction represents a quantification of that judgment. A meteorologist's predictions may not always correspond to his judgments (refer to Section 2).

[5] The *evaluation* problem in meteorology is, of course, a problem with which many meteorologists are familiar. We shall consider this problem from a general point of view which, we believe, permits the formulation of a *comprehensive* framework.

[6] The term "assessment" is equivalent to the terms "prediction" and "statement."

"honest" assessments, 2) to evaluate meteorologists, i.e., to identify "good" assessors, and 3) to help meteorologists to become "better" assessors. Section 5 contains a brief summary.

2. Kinds of "goodness"

a. The notion of "goodness"

de Finetti (1962), in the title of a recent article, asks, "Does it make sense to speak of 'good probability appraisers'?" We believe that this question can be answered in the affirmative, since we are, in general, concerned with evaluating assessors in a particular context, in this paper in a meteorological context.[7] We shall attempt to intuitively justify this belief by a heuristic argument.

Suppose that we, the evaluators, must select an individual to assess the probability of a particular meteorological event, such as the probability of measurable precipitation at Hartford, Conn., tomorrow. Clearly, we will seriously consider, for this task, only a small subset of the set of all possible assessors. However, the mere selection of a small set of potential assessors implies that the members of this set will, in general, be "better" assessors than the nonmembers. Thus, we can speak of the "goodness" of assessors, at least in two "degrees." In a similar manner, we may be able to select a subset of the set of potential assessors, which, in turn, indicates the existence of finer degrees of "goodness." Note that this process generates successively finer partitions of the set of all possible assessors. Each partition is, in general, more difficult to realize than the preceding partition simply because finer degrees of "goodness" are more difficult to determine. In Section 3 we consider several measures which can assist in determining such degrees of "goodness." However, we must first describe more precisely certain aspects of the notion of "goodness" upon which the measures of "goodness" are based.

Two general *standards* of "goodness," standards which could serve as the basis for the selection of an assessor, suggest themselves. First, since the task involves the assessment of probabilities, we would like an assessor to have some understanding of the concept of probability and some aptitude for dealing with numbers. Second, we would like an assessor to have a general knowledge of meteorology and, in addition, some specific knowledge of the problem of concern (e.g., precipitation at Hartford). We shall refer to these standards as the *normative* and *substantive* standards of "goodness," respectively.

[7] From the point of view of most meteorologists, the only appropriate answer to this question is an affirmative answer. However, from the point of view of an individual who subscribes to the theory of subjective probability, an affirmative answer may not be appropriate.

b. Normative "goodness"

The normative standard of "goodness" relates to the theory of subjective probability. Subjective probabilities assessed in accordance with certain plausible postulates of coherence can be shown to conform mathematically to a probability measure (de Finetti, 1937; Savage, 1954). In essence, these postulates are such that a series of bets cannot be set up against an individual who obeys the postulates in such a manner that the individual is certain to lose regardless of the outcome of the event(s) of concern. The theory of subjective probability prescribes, in addition, that an individual's assessments should correspond to his judgments, since the assessments represent a quantification of these judgments. Thus, a "perfect" assessor, in a normative sense, obeys the postulates of coherence and makes assessments that correspond completely to his judgments.

c. Substantive "goodness"

The substantive standard of "goodness" relates to the context in which the assessments are made. In a particular context, a "good" assessor is an assessor who is extremely knowledgeable with respect to the problem under consideration. We shall assume that an assessor's knowledge manifests itself in predictions which, to a greater or lesser degree, possess certain "desirable" properties. In order to identify these properties we must first consider the reasons why the assessor, a meteorologist in this paper, prepares probabilistic predictions.

Meteorologists prepare predictions 1) to provide a means of ascertaining (through the process of evaluation) the extent of their knowledge (as expressed by their predictions) and 2) to provide the user of meteorological predictions, i.e., the "decision maker," with information upon which to base his decisions. Meteorologists express their predictions in *probabilistic* terms 1) since their knowledge contains, in general, an element of uncertainty (e.g., Sutcliffe, 1954; Malone, 1957) and 2) since, in the presence of this uncertainty, probabilistic predictions are of greater "value" to decision makers than categorical predictions (e.g., Thompson, 1952; Nelson and Winter, 1964). We shall refer to these reasons for the preparation of probabilistic predictions as the *meteorological* and *operational* reasons, respectively.

From a meteorological point of view, an evaluator is concerned with determining the extent of a meteorologist's knowledge as expressed by his predictions. If the meteorologist's knowledge is complete, then his predictions would (presumably) be "perfect," i.e., categorical and correct. However, since the meteorologist's knowledge is, in general, incomplete, the evaluator is concerned with the "degree of perfection" of the predictions, i.e., with the "degree of association" between the predictions and the relevant observations. The

process of determining the degree of association between predictions and observations has been referred to as *empirical evaluation* (Murphy and Epstein, 1967a; Murphy, 1968b). The properties, or attributes, of predictions which are "desirable" from the point of view of empirical evaluation depend, of course, upon particular definitions of the term "association." The term "association," in a meteorological context, appears to possess two particularly relevant definitions (Murphy and Epstein, 1967a; Murphy, 1968b). These definitions have led to the identification of two attributes: 1) the attribute *validity* based upon the notion of association between predictions and observations on an individual basis and 2) the attribute *bias* based upon the notion of association between predictions and observations on a collective basis. In this paper we shall be concerned only with the attribute validity.[8] A prediction which is categorical and correct is completely valid, while a prediction which is categorical and incorrect is completely invalid. Thus, a probabilistic prediction is, in general, partially valid.

From an operational point of view, an evaluator is concerned with determining the "value" of a meteorologist's predictions to decision makers. In the context of the standard subjective framework for decision making (e.g., Fishburn, 1964), the value of a "consequence" to a decision maker is expressed in terms of its "utility" [in the sense of von Neumann and Morgenstern (1953)]. Thus, the attribute of the predictions which is of concern from an operational point of view is their *utility*. The process of determining the utility of predictions has been referred to as *operational evaluation* (Murphy and Epstein, 1967a; Murphy, 1968b). In this paper we shall be primarily concerned with those aspects of substantive "goodness" which relate to empirical evaluation.

d. Normative and substantive "goodness"

In essence, then, the normative standard of "goodness" concerns expertise in probability assessment, while the substantive standard of "goodness" concerns expertise in the domain in which the assessments are made. The former requires probabilities to correspond to judgments, while the latter requires probabilities to correspond to something in reality. We, as evaluators, would, of course, select that assessor who (completely) meets both of these standards. However, we will seldom encounter any such assessor in practice. Further, since the standards require expertise in two different domains, we may often find that an assessor does not (adequately) meet one of the two standards. For example, an individual, e. g., a statistician, who is able to make coherent assessments in accordance with his judgments, may know very little about meteorology. A meteorologist, on the other hand, who is an expert in his field, may not understand the concept of subjective probability. Of course, we do not have to choose between two such assessors, since meteorologists, with some instruction and practice, can learn to make assessments in accordance with the normative standards of "goodness" [refer to Winkler (1967a)][9].

3. Measures of "goodness"

a. The notion of scoring rules

The two standards of "goodness" provide a basis upon which to select a relatively small set of potential assessors. However, we may assume that, in general, a number of assessors will "meet" these general standards. Suppose that we want to identify, if possible, the "best" assessor.[10] Then, what we require, of course, are *measures* of "goodness" which determine the degree to which each assessor meets these standards. Such measures will, of course, be concerned with particular attributes of the predictions for each standard.

From a normative as well as a substantive point of view, we are concerned with the association between the predictions and the relevant observations.[11] Thus, a measure of "goodness" will be a function of an assessor's probabilities and the relevant observation. We shall refer to such a measure as a *scoring rule*.

b. Scoring rules and normative "goodness": Proper scoring rules

1) THE NOTION. Consider an individual who must make a probability assessment for an event E which consists of n mutually exclusive and collectively exhaustive outcomes $E_1, \cdots, E_n$. If the outcome of an event is a continuous variable on the interval (a,b), we shall assume that the $E_i (i=1, \cdots, n)$ form an n-fold partition on (a,b). Let the row vector $\mathbf{r}=(r_1, \cdots, r_n)$ denote the individual's assessment, where z_i is the assessor's stated probability that E_i will occur. Suppose that the row vector $\mathbf{p}=(p_1, \cdots, p_n)$ represents the assessor's true judgment, where p_i is his subjective probability that E_i will occur. Further, let the row

[8] The attribute validity is the attribute of primary concern for empirical evaluation (Murphy, 1968b).

[9] Some meteorologists have, of course, already become quite proficient at probability assessment (e.g., Root, 1962; Sanders, 1963; Hughes, 1965). However, a better understanding of the concept of subjective probability and its implications for assessing and evaluating probabilities would hopefully make meteorologists even more proficient (cf. Epstein, 1966; refer to Section 4).

[10] The question of whether the identification of the "best" assessor is an appropriate and/or a realizable objective is briefly considered in Section 4 [refer also to Murphy and Epstein (1967a)].

[11] This statement holds whether we are concerned with empirical evaluation or operational evaluation (Murphy, 1968b).

vector $\mathbf{d}=(d_1, \cdots, d_n)$ represent the observation, where d_i equals one if E_i occurs and zero otherwise.

We have stated, in Section 2, that a "perfect" assessor, in the normative sense, is an assessor who obeys the postulates of coherence and makes assessments that correspond (completely) to his judgments. The first condition imposes certain constraints on the probabilities, namely, $r_i \geq 0$, $p_i \geq 0$, $\sum_i r_i = 1$ and $\sum_i p_i = 1$ $(i=1, \cdots, n)$. An assessor who violates these constraints and is informed of his violation can, of course, revise his probabilities (in accordance with these postulates). The second condition requires that $\mathbf{r}=\mathbf{p}$. Of course, since we do not know $\mathbf{p}$, we cannot compare $\mathbf{r}$ and $\mathbf{p}$. However, the scoring rules which we shall consider are such that an assessor can maximize his expected score if and only if he sets $\mathbf{r}$ equal to $\mathbf{p}$. A scoring rule which possesses this property will be referred to as a *proper* scoring rule.[12] Since we are concerned with the association between a prediction and an observation, a proper scoring rule is a function of both $\mathbf{r}$ and $\mathbf{d}$.

2) EXAMPLES. We consider, in this paper, three proper scoring rules: (i) the quadratic scoring rule, (ii) the spherical scoring rule and (iii) the logarithmic scoring rule.[13]

(i) Quadratic scoring rule. The quadratic scoring rule $Q(\mathbf{r}, \mathbf{d})$ is defined as (a prime denoting a column vector)

$$Q(\mathbf{r}, \mathbf{d})=[1-(\mathbf{r}-\mathbf{d})(\mathbf{r}-\mathbf{d})'],$$

or

$$Q(\mathbf{r},\mathbf{d})=[1-\sum_i (r_i-d_i)^2].$$

Note that the range of $Q(\mathbf{r},\mathbf{d})$ is $[-1, 1]$. If outcome E_j occurs, $d_j=1$ and $d_i=0$ for all $i \neq j$. Then, $Q(\mathbf{r}, \mathbf{d})$ equals $Q_j(\mathbf{r}, \mathbf{d})$, where

$$Q_j(\mathbf{r},\mathbf{d})=(2r_j-\sum_i r_i^2). \tag{1}$$

The assessor's expected score is $E[Q(\mathbf{r}, \mathbf{d})]$, or simply $E(Q)$, where

$$E(Q)=\sum_j p_j Q_j(\mathbf{r},\mathbf{d}),$$

or, from (1),

$$E(Q)=\sum_j p_j(2r_j-\sum_i r_i^2),$$

or

$$E(Q)=\sum_j p_j^2-\sum_j (r_j-p_j)^2. \tag{2}$$

Note, in (2), that the assessor, to maximize his expected score, must set $\mathbf{r}$ equal to $\mathbf{p}$. Thus, the quadratic scoring rule is proper. Note that the assessor's expected score when he sets $\mathbf{r}$ equal to $\mathbf{p}$ is, from (2), $\sum_j p_j^2$.

(ii) Spherical scoring rule. The spherical scoring rule $S(\mathbf{r}, \mathbf{d})$ is defined as

$$S(\mathbf{r}, \mathbf{d})=\mathbf{d}\mathbf{r}'/(\mathbf{r}\mathbf{r}')^{\frac{1}{2}},$$

or

$$S(\mathbf{r},\mathbf{d})=\sum_i r_i d_i/(\sum_i r_i^2)^{\frac{1}{2}}.$$

Note that the range of $S(\mathbf{r}, \mathbf{d})$ is $[0, 1]$. If outcome E_j occurs, $S(\mathbf{r}, \mathbf{d})=S_j(\mathbf{r}, \mathbf{d})$, where

$$S_j(\mathbf{r},\mathbf{d})=r_j/(\sum_i r_i^2)^{\frac{1}{2}}.$$

Then, the assessor's expected score is $E(S)$, where

$$E(S)=\sum_j p_j r_j/(\sum_i r_i^2)^{\frac{1}{2}},$$

or

$$E(S)=\sum_j p_j r_j/(\sum_j r_j^2)^{\frac{1}{2}}. \tag{3}$$

From Schwarz's inequality (e.g., Halmos, 1958),

$$\sum_j p_j r_j \leq (\sum_j r_j^2)^{\frac{1}{2}}(\sum_j p_j^2)^{\frac{1}{2}}.$$

Thus,

$$E(S) \leq (\sum_j p_j^2)^{\frac{1}{2}},$$

with equality holding if and only if $r_j=kp_j$ for all j (k is a constant). The assessor's expected score is maximized, of course, if equality holds. Note that, since $\sum_j r_j=\sum_j p_j=1$, k equals one and the spherical scoring rule is proper. When the assessor sets $\mathbf{r}$ equal to $\mathbf{p}$ his expected score is, from (3), $(\sum_j p_j^2)^{\frac{1}{2}}$.

(iii) Logarithmic scoring rule. The logarithmic scoring rule $L(\mathbf{r}, \mathbf{d})$ is defined as

$$L(\mathbf{r}, \mathbf{d})=\ln(\mathbf{d}\mathbf{r}'),$$

or

$$L(\mathbf{r},\mathbf{d})=\ln(\sum_i d_i r_i).$$

Note that the range of $L(\mathbf{r},\mathbf{d})$ is $(-\infty, 0]$. If outcome E_j occurs,

$$L_j(\mathbf{r},\mathbf{d})=\ln r_j.$$

---

[12] We have implicitly assumed that the scoring rule of concern is defined in such a manner that a larger score is "better." Such a rule may be said to have a positive orientation. However, if a scoring rule is defined in such a manner that a smaller score is "better" (i.e., if the rule has a negative orientation), this rule is also proper if an assessor must set $\mathbf{r}$ equal to $\mathbf{p}$ in order to *minimize* his expected score.

[13] For a general discussion of scoring rules refer to Winkler (1967b). For a discussion of proper and improper scoring rules, in a meteorological context, refer to Murphy and Epstein (1967b) and Murphy (1968b).

TABLE 1. Proper scoring rules.

Rule	Original Form	Original Range	Standard Form	Standard Range
Quadratic $Q(\mathbf{r},\mathbf{d})$	$1-\sum_i (r_i-d_i)^2$	$[-1, 1]$	$1-(1/2)\sum_i (r_i-d_i)^2$	$[0,1]$
Spherical $S(\mathbf{r},\mathbf{d})$	$\sum_i r_i d_i/(\sum_i r_i^2)^{\frac{1}{2}}$	$[0,1]$	$\sum_i r_i d_i/(\sum_i r_i^2)^{\frac{1}{2}}$	$[0,1]$
Logarithmic $L(\mathbf{r},\mathbf{d})$	$\ln(\sum_i d_i r_i)$	$(-\infty, 0]$	$1+\ln(\sum_i d_i r_i)$	$(-\infty, 1]$

Then, the assessor's expected score is $E(L)$,where

$$E(L)=\sum_j p_j \ln r_j. \tag{4}$$

Maximizing $E(L)$ in (4) is equivalent to maximizing

$$E(L)-\lambda(\sum_j r_j-1)=\sum_j p_j \ln r_j-\lambda(\sum_j r_j-1), \tag{5}$$

since $\sum_j r_j=1$ (λ is a Lagrange multiplier). Differentiating (5) with respect to r_i and setting the result equal to zero yields

$$r_i=(1/\lambda)p_i.$$

Since $\sum_i r_i=\sum_i p_i=1$, λ equals one and the logarithmic scoring rule is proper. Note that when $\mathbf{r}$ equals $\mathbf{p}$, $E(L)$ in (4) equals $\sum_j p_j \ln p_j$, which is equivalent to Shannon's information measure (Shannon and Weaver, 1949).

(iv) Other scoring rules. The quadratic, spherical and logarithmic scoring rules are *not* the only proper scoring rules.[14] In fact, a linear combination of proper rules is a proper rule.[15] In particular, a linear function of a proper rule is a proper rule. That is, if $M(\mathbf{r},\mathbf{d})$ is a proper rule, then $M^*(\mathbf{r},\mathbf{d})$, where

$$M^*(\mathbf{r},\mathbf{d})=a[M(\mathbf{r},\mathbf{d})]+b, \tag{6}$$

in which a and b are constants, is a proper rule.[16] The linear transformation in (6) simply effects a change in scale. That is, if the range of $M(\mathbf{r},\mathbf{d})$ is (x,y), then the range of $M^*(\mathbf{r},\mathbf{d})$ will be $(ax+b, ay+b)$. Thus, we can, if we wish, establish a "standard" range for proper scoring rules (at least for those proper rules with a finite range). The (closed) unit interval $[0,1]$, where larger scores correspond to "better" predictions, is such a standard range. The original and standard forms of the quadratic, spherical and logarithmic scoring rules are indicated in Table 1. Note that, since the range of the (original) logarithmic rule is infinite, the range of the standard logarithmic rule is infinite.[17]

The probability score PS (Brier, 1950) is a linear function of the quadratic rule, In particular,

$$PS=1-Q(\mathbf{r},\mathbf{d}).$$

The range of the probability score is $[0,2]$. The validity measure v (Murphy, 1968b) is also a linear function of the quadratic rule.[18] In particular,

$$v=(\tfrac{1}{2})[1+Q(\mathbf{r},\mathbf{d})].$$

Thus, the range of the validity measure is $[0,1]$. Of course, both the probability score PS and the validity measure v are proper scoring rules. For a discussion of proper and improper scoring rules, in a meteorological context, refer to Murphy and Epstein (1967b) and Murphy (1968b).

c. Scoring rules and substantive "goodness"

Scoring rules are, by definition, measures of the association between the predictions and the relevant observations. Proper scoring rules, in particular, are measures of the association between predictions and observations on an individual basis. Thus, proper scoring rules are measures of the attribute validity. However, all proper scoring rules are not concerned with the same aspects of this attribute. For example, the logarithmic scoring rule is concerned only with the probability of the outcome that occurs. The quadratic and spherical scoring rules, on the other hand, are concerned with all the probabilities, each in a somewhat different manner. Thus, we can state, in general terms, that the quadratic and spherical scoring rules are "total" measures of validity, while the logarithmic scoring rule is only a "partial" measure of validity. However, the differences between these proper scoring rules cannot, at this time,

[14] The class of proper scoring rules has been described by de Finetti and Savage (1963). For a bibliography on scoring rules refer to Murphy and Winkler (1968).

[15] More precisely, a linear combination of proper scoring rules is proper if (a) all the scoring rules of concern have the same orientation *and* (b) all the coefficients (in the linear combination) have the same sign.

[16] Note that if a is positive both $M(\mathbf{r},\mathbf{d})$ and $M^*(\mathbf{r},\mathbf{d})$ have the same orientation while if a is negative $M(\mathbf{r},\mathbf{d})$ and $M^*(\mathbf{r},\mathbf{d})$ have the opposite orientation.

[17] Note, from (6), that the logarithms in the logarithmic rule can be taken with respect to any base.

[18] The validity measure v is the square of the distance measure V described by Epstein and Murphy (1965).

TABLE 2. Meteorologists' assessments.

Meteorologists	Assessments r_1	r_2	r_3
A	0.35	0.60	0.05
B	0.30	0.35	0.35

TABLE 3. Meteorologists' scores.

Meteorologists	Scoring rules $Q(\mathbf{r},\mathbf{d})$	$S(\mathbf{r},\mathbf{d})$	$L(\mathbf{r},\mathbf{d})$
A	0.215	0.503	−1.050
B	0.265	0.518	−1.204

be explained satisfactorily in terms of a concern with different aspects of the attribute validity. We shall briefly consider certain related aspects of substantive "goodness" in Section 4. For a discussion of substantive "goodness" in a meteorological context refer to Murphy and Epstein (1967a) and Murphy (1968a).

4. Uses of measures of "goodness"

We shall briefly consider in this section three uses of our measures of "goodness," the proper scoring rules: 1) to encourage assessors to be "honest," 2) to evaluate, i.e., to determine the "goodness" of, assessors and 3) to help individuals to become "better" assessors.

The use of proper scoring rules will encourage an assessor to be "honest," i.e., to make his assessment **r** correspond to his judgment **p**, since a proper scoring rule is such that the assessor maximizes his expected score when he sets **r** equal to **p**. Thus, proper scoring rules should play an important role in the assessment of subjective probabilities in meteorology. The use of such rules will encourage meteorologists to state their true beliefs rather than to "hedge." In fact, the use of proper scoring rules ensures that a meteorologist's "best" hedge is *no* hedge [refer to Murphy and Epstein (1967b)].

As indicated in Section 3, proper scoring rules provide a means of determining the "goodness" of assessors from a substantive, as well as a normative, point of view.[19] Thus, we can evaluate assessors simply by computing their scores. Further, we can determine the relative "goodness" of assessors by comparing their scores. Such a comparison would, in general, be based upon a collection of predictions. Specifically, we could simply rank the assessors according to their *average* scores.[20] However, different proper scoring rules may yield different rankings of the assessors. Consider, for example, two meteorologists A and B whose assessments of the probability of an event E, with outcomes E_1, E_2 and E_3, on a particular occasion are indicated in Table 2. Suppose that on this occasion outcome E_1 occurs. Then, according to the three proper scoring rules considered in Section 3, which meteorologist is a "better" assessor on this occasion? The meteorologists' scores are indicated in Table 3. Note that the quadratic and spherical scoring rules indicate that B is a "better" assessor, while the logarithmic scoring rule indicates that A is a "better" assessor. Thus, proper scoring rules may not yield consistent results.[21] More specifically, the answer to the question of which meteorologist is a "better" assessor may depend upon the particular aspect of the attribute validity with which the evaluator is concerned.[22] However, we do have evidence that rankings based upon *average* scores will be reasonably consistent (Winkler and Murphy, 1968).

Another use of proper scoring rules is to help individuals to become "better" assessors. Just as an examination and the resulting score enable a student to determine his strong and weak points and to compare himself with his fellow students, the process of assessing the probabilities and the resulting score should enable an assessor to evaluate his own performance. Thus, the results obtained by scoring the assessments, when used as feedback, could serve as a learning device. Specifically, experience should help an assessor to understand the correspondence between judgments and probabilities and to confirm the fact that he should set **r** equal to **p** in order to maximize his expected score. Further, interpersonal comparisons of scores should lead those assessors with "poorer" scores to reevaluate the mental processes involved in their assessments. For example, a meteorologist whose scores are consistently "poor" because of his tendency to overestimate the probability of precipitation should reexamine both his judgments (concerning, for example, the meteorological conditions relevant to the prediction of precipitation) and the rela-

[19] This statement should not be taken to mean that these scoring rules are *objective* measures of "goodness" or even that such objective measures exist. The "goodness" of an assessor should be thought of as a subjective judgment of another individual, e.g., the evaluator, concerning the assessor. To the extent that this individual believes that the scores are related to factors such as skill or expertise, he will believe that the scores provide a rough measure of "goodness."

[20] As an example of the application of such a procedure in a field other than meteorology, de Finetti (1965) and Shuford *et al.* (1966) have proposed that students answer examination questions in probabilistic terms and that proper scoring rules be used to evaluate the students.

[21] Specifically, proper scoring rules which are not linearly related will, in general, yield inconsistent results for certain assessments except in the situation in which the event of concern consists of only two outcomes [refer to Murphy (1968b)].

[22] For example, a Bayesian model has been developed in which likelihood ratios are used to compare the probabilistic predictions of a number of assessors (H. V. Roberts, 1965, 1968). If $n>2$, the only proper scoring rule which is consistent with this particular model is the logarithmic scoring rule (Winkler, 1968). Of course, other scoring rules may be consistent with different models.

tionship between these judgments and the assessments. In this way, an individual, e.g., a meteorologist, can improve his performance with respect to both the normative and substantive standards of "goodness" through the use of proper scoring rules. For a general discussion of learning by experience in probability assessment refer to Winkler (1967a, b).

5. Summary

The probabilities that constitute probabilistic predictions in meteorology are, in general, subjective probabilities. Thus, when we consider the evaluation problem, i.e., the problem of determining the "goodness" of the predictions, or of the meteorologists who issue the predictions, we must consider two kinds of "goodness": 1) normative "goodness" and 2) substantive "goodness." In essence, the normative standard of "goodness" concerns expertise in probability assessment, while the substantive standard of "goodness," in this context, concerns expertise in meteorology. Thus, measures of "goodness" must, of course, concern themselves with both standards of "goodness."

We have described, in this paper, several measures, or scoring rules, which, by construction, encourage meteorologists to meet, in full, the normative standard of "goodness." In particular, such rules encourage a meteorologist to make his assessments correspond to his judgments. We referred to such measures as *proper* scoring rules. These rules also serve as measures of substantive "goodness." Specifically, proper scoring rules are measures of (some aspect of)the attribute validity. However, since proper rules may not yield consistent results, the nature of, and the relationship between, such rules, as measures of substantive "goodness," require further study.

We have, in addition, briefly considered several uses of such rules: 1) to encourage meteorologists to be "honest," 2) to evaluate meteorologists (the purpose which such measures usually serve), and 3) to help meteorologists to become "better" assessors.

Finally, the realization that a meteorologist's predictions should meet certain normative, as well as substantive, standards of "goodness" has provided, we believe, a sound basis upon which to develop a comprehensive framework for evaluating probabilistic predictions in meteorology. In particular, the normative standard of "goodness" has suggested one property which all measures of "goodness" should possess; namely, the property of encouraging a meteorologist to make his assessments correspond to his judgments. If we want to identify the "best" assessor, we must, then, search for other "desirable" properties.

REFERENCES

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

de Finetti, B., 1937: La prevision: Ses lois logiques, ses sources subjectives. *Ann. Inst. Poincare*, **7**, 1–68. Translated as Foresight: Its logical laws, its subjective sources in *Studies in Subjective Probability*, New York, John Wiley and Sons, 1964, 203 pp.

——, 1962: Does it make sense to speak of 'good probability appraisers'? *The Scientist Speculates: An Anthology of Partly-Baked Ideas*, New York, Basic Books, 413 pp.

——, 1965: Methods for discriminating levels of partial knowledge concerning a test item. *Brit. J. Math. Statis. Psychol.*, **18**, 87–123.

——, and L. J. Savage, 1963: The elicitation of subjective probabilities. University of Rome, and University of Michigan, unpublished manuscript, 36 pp.

Enger, I., L. J. Reed and J. E. MacMonegle, 1962: An evaluation of 2-7 hour aviation terminal-forecasting techniques. Hartford, Conn., Travelers Research Center, Inc., Contract FAA/BRD-363, Tech. Publ. 20, 27 pp.

Epstein, E. S., 1966: Quality control for probability forecasts. *Mon. Wea. Rev.*, **94**, 487–494.

——, and A. H. Murphy, 1965: A note on the attributes of probabilistic predictions and the probability score. *J. Appl. Meteor.*, **4**, 297–299.

Federal Aviation Agency/U. S. Weather Bureau, 1965: Test plan: Implementation of REEP for forecasting ceiling and visibility. Washington, D. C., Federal Aviation Agency (Air Traffic Control Procedures Branch) and U. S. Weather Bureau (Techniques Development Laboratory), FAA Project 150-535-03A, manuscript, 15 pp.

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

Halmos, P. R., 1958: *Finite-Dimensional Vector Spaces*. Princeton, N. J., D. Van Nostrand Co., 200 pp.

Hughes, L. A., 1965: On the probability forecasting of the occurrence of precipitation. Washington, D. C., Dept. of Commerce, ESSA, Weather Bureau, Tech. Note 20-CR-3, 36 pp.

Malone, T. F., 1957: Applied meteorology. *Meteorological Research Reviews: Summaries of Progress from 1951 to 1955. Meteor. Monogr.*, **3**, Nos. 12-20, 152–159.

Murphy, A. H., 1968a: Probabilistic predictions in meteorology: A bibliography. Washington, D. C., Dept. of Commerce, ESSA, Weather Bureau, Techniques Development Laboratory, Tech. Memo., in preparation.

——, 1968b: The evaluation of probabilistic predictions in meteorology. Ann Arbor, University of Michigan, Dept. of Meteorology and Oceanography, Tech. Rept., Contract Cwb-10847 and NSF Grant GA-906.

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

——, and ——, 1967b: A note on probability forecasts and "hedging." *J. Appl. Meteor.*, **6**, 1002–1004.

——, and R. L. Winkler, 1968: Scoring rules: A bibliography. Hartford, Conn., Travelers Research Center, Inc., manuscript, in preparation.

Nelson, R. R., and S. G. Winter, 1964: A case study in the economics of information and coordination: The weather forecasting system. *Quart. J. Econ.*, **78**, 420–441.

Roberts, C. F., 1965: On the use of probability statements in weather forecasts. Washington, D. C., Dept. of Commerce, ESSA, Weather Bureau, Tech. Note 8-FCST-1, 15 pp.

Roberts, H. V., 1965: Probabilistic prediction. *J. Amer. Statis. Assoc.*, **60**, 50–62.

——, 1968: On the meaning of the probability of rain. *Proc. First Nat. Conf. Statis. Meteor.* Boston, Amer. Meteor. Soc., 133–141.

Root, H. E., 1962: Probability statements in weather forecasting. *J. Appl. Meteor.*, **1**, 163–168.

Sanders, F., 1963: On subjective probability forecasting. *J. Appl. Meteor.*, **2**, 191–201.

Savage, L. J., 1954: *The Foundations of Statistics*. New York, John Wiley and Sons, 294 pp.

Shannon, C. E., and W. Weaver, 1949: *The Mathematical Theory of Communication*. Urbana, University of Illinois Press, 117 pp.

Shuford, E. H., A. Albert and H. E. Massengill, 1966: Admissible probability measurement procedures. *Psychometrika*, **31**, 125–145.

Sutcliffe, R. C., 1954: Predictability in meteorology. *Arch. Meteor. Geophys. Bioklim.*, **A7**, 3–15.

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

von Neumann, J., and O. Morgenstern, 1953: *Theory of Games and Economic Behavior*. Princeton University Press, 641 pp.

Winkler, R. L., 1967a: The assessment of prior distributions in Bayesian analysis. *J. Amer. Statis. Assoc.*, **62**, 776–800.

——, 1967b: The quantification of judgment: Some methodological suggestions. *J. Amer. Statis. Assoc.*, **62**, 1105–1120.

——, 1968: Scoring rules and the evaluation of probability assessors. Bloomington, Indiana University, manuscript, in preparation.

——, and A. H. Murphy, 1968: Evaluation of subjective precipitation probability forecasts. *Proc. First Nat. Conf. Statis. Meteor.*, Boston, Amer. Meteor. Soc., 148–157.