Acta Psychologica 34 *Subjective probability (G. de Zeeuw et al., eds.)* 1970, 273–286

SCORING RULES IN PROBABILITY ASSESSMENT AND EVALUATION*

ALLAN H. MURPHY
University of Michigan, Ann Arbor, Mich. 48104, U.S.A.

and

ROBERT L. WINKLER**
Indiana University, Bloomington, Ind. 47101, U.S.A.

ABSTRACT

The purpose of this paper is to briefly discuss some important current questions and problems related to the use of scoring rules (SRs) both in connection with the actual assessment of probabilities and with the evaluation of probability forecasts and probability assessors. With regard to the *assessment* process, we consider both the case in which the assessor's utility function is linear and the case in which his utility function is nonlinear. Under linear utility, important problems of concern are the sensitivity of SRs to deviations from optimality (with a strictly proper SR, optimality consists of the assessor making his statements correspond to his judgments) and the effect of psychological considerations arising from the use of different SRs. Under nonlinear utility, SRs should be modified to allow for the nonlinearity in such a manner that for a specific utility function, the modified SRs are strictly proper. This introduces the difficult question of the assessment of the assessor's utility function. With regard to the *evaluation* process (as opposed to the assessment process), we consider the process from an inferential viewpoint and from a decision-theoretic viewpoint. From an inferential viewpoint, attributes such as validity may be of interest, and in certain circumstances these attributes may be related to SRs. The attributes of interest, of course, depend on the framework within which the evaluation process is undertaken. From a decision-theoretic viewpoint, SRs may be related to a decision maker's utilities or expected utilities (under uncertainty about the utilities) if the decision maker uses the assessed probabilities in an actual decision situation.

In summary, there are many important questions and problems related to SRs, and the need for future research on these problems seems clear. Such research should lead to a greatly improved understanding of the processes of probability assessment and evaluation.

1. INTRODUCTION

The personalistic theory of probability prescribes that the probabilities to be used in inferential and decision-making situations should correspond

* Supported, in part, by the National Science Foundation (Atmospheric Sciences Section) under Grant GA-1707.

** Presently on leave and visiting at the University of Washington, Seattle, Wash. 98105, U.S.A.

with the assessor's judgments. Since the judgments exist solely in the assessor's mind, there is no way to determine whether or not this requirement is satisfied. However, by rewarding or penalizing the assessor according to certain scoring rules (SRs), one can encourage an assessor to make his stated probabilities correspond with his judgments. SRs, which involve the computation of a score based on the assessor's stated probabilities and on the event which actually occurs, are useful in the evaluation of probability assessors as well as in the assessment process itself.

General discussions involving SRs and reviews of the previous work in the area may be found in WINKLER (1967), WINKLER and MURPHY (1968), DE FINETTI and SAVAGE (1969), and STAËL VON HOLSTEIN (1970). The reader interested in the historical development and use of the concept of SRs should consult these sources and the references cited there. The purpose of this paper is to briefly discuss some important current questions and problems related to the use of SRs, both in connection with the actual assessment of probabilities and with the evaluation of probability forecasts and probability assessors. In section 2, the concepts of assessment and evaluation are briefly described and compared. Some problems which are of particular concern with regard to the assessment process are discussed in section 3, and some problems which are of particular concern with regard to the evaluation process are discussed in section 4.

2. ASSESSMENT AND EVALUATION

The role of SRs in probability *assessment* is to encourage the assessor to be 'honest', i.e., to make his statements correspond to his judgments. Thus, SRs which encourage honesty on the part of the assessor, i.e., 'proper' SRs (refer to section 3), are of primary interest. Further, since assessment is an a priori task (a task which takes place in the absence of complete knowledge of the 'true' state), *expected*, rather than actual, scores are of primary interest. However, the actual scores are of some (secondary) interest, since the assessor's actual scores may influence his behavior.

The role of SRs in probability evaluation is to evaluate, i.e., to measure the (substantive) 'goodness' of, the probabilities. In this task the SRs need not necessarily be 'proper' SRs (however, see below). Since evaluation is an a posteriori task (a task which takes place in the presence of complete

knowledge of the true state), *actual*, rather than expected, scores are of primary interest.

In the previous paragraphs the assessment and evaluation tasks have been considered separately. However, as WINKLER (1969) has indicated, game theoretic problems may arise if the assessor is rewarded (or penalized) and evaluated with different SRs. These problems can be eliminated if the same SR is utilized in *both* the assessment and the evaluation tasks. Since 'proper' SRs are of primary interest for the assessment task, we restrict our attention in this paper largely to such SRs.

3. THE ASSESSMENT PROCESS

The primary role of SRs with respect to the assessment process (the actual process of quantifying one's judgments and expressing them in terms of probabilities) is to encourage the assessor to be honest in reporting his true judgments and to take the task of assessment seriously and devote considerable time and care to the assessment process. In our discussion of the assessment process, it will be convenient to consider two cases: (1) the situation in which the assessor's utility function for the score is linear, and (2) the situation in which his utility function is nonlinear.

3.1. *Linear utility*

A scoring rule provides the assessor with a 'payoff' which depends on his stated probability assessments and on the event which actually occurs. In decision theory, the axioms of rational choice imply that a person's judgments about uncertain situations can be represented by subjective, or personal, probabilities; that a person's preferences for various consequences can be represented by a utility function; and that in choosing among alternative actions, a person should choose the action which maximizes his expected utility (e.g., see SAVAGE, 1954; or FISHBURN, 1964). But if the 'consequence' to an assessor is a linear function of some SR, and the assessor's utility function for the score is linear, then maximization of the expected score is equivalent to maximization of expected utility. [1]

[1] We have implicitly assumed that the SR 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 specific SR is defined in such a manner that a smaller score is 'better' (i.e., if the rule has a negative orientation), then the assessor should attempt to minimize his expected score.

Suppose that an assessor must make a probability forecast in a situation in which there are n mutually exclusive and collectively exhaustive outcomes, $E_1, E_2, \ldots, E_n$. Let r_i denote the assessor's probability forecast for the outcome E_i, and let p_i denote the assessor's true judgment regarding the probability that E_i will occur. Of course, $p_i \geq 0$ and $r_i \geq 0$ for $i = 1, 2, \ldots, n$, and $\sum p_i = \sum r_i = 1$. To simplify the notation, let $r = (r_1, r_2, \ldots, r_n)$ and let $p = (p_1, p_2, \ldots, p_n)$. Then the *expected score* for a given SR is

$$\mathrm{SR}(r,p) = \sum_{h=1}^{n} p_h \, \mathrm{SR}_h(r),$$

where the subscript h refers to the event which actually occurs and $\mathrm{SR}_h(r)$ is the score corresponding to a stated r when E_h occurs. SR is said to be *strictly proper* if $\mathrm{SR}(p, p) > \mathrm{SR}(r, p)$ for all $r \neq p$. If the inequality is not strict, then SR is said to be *proper*. A strictly proper SR obliges the assessor to set r equal to p in order to maximize his expected score; with a proper (but not strictly proper) SR, setting $r = p$ will maximize the expected score, but other choices of r may also enable the assessor to attain the maximum expected score.

With regard to the assessment process, it seems desirable to limit the choice of SRs to strictly proper SRs if possible, or at least to proper SRs. There are many such rules, so this should present no problem. The difficulty arises when one attempts to compare the various strictly proper (or proper) SRs. If they all satisfy the criteria of keeping the assessor honest and encouraging careful assessment, which one should be used in any actual probability forecasting situation?

One question which has not been investigated in detail is the question of the sensitivity of the expected scores to deviations from one's true judgments. This, of course, depends on p and on the particular rule used. It would seem that a 'sharper' (i.e., more sensitive) SR would be more likely to encourage careful assessment than a 'flatter' rule, because deviations from optimality are more costly with a sensitive rule than with a relatively insensitive one. When $n = 2$ (the two-state situation), a comparison was made among three strictly proper SRs, the logarithmic, quadratic, and spherical SRs, which are defined for any n as follows:

$$L_h(r) = \log r_h, \quad Q_h(r) = (1 + 2r_h - \sum_{i=1}^{n} r_i^2)/2, \quad \text{and} \quad S_h(r) = r_h / \sqrt{\sum_{i=1}^{n} r_i^2}.$$

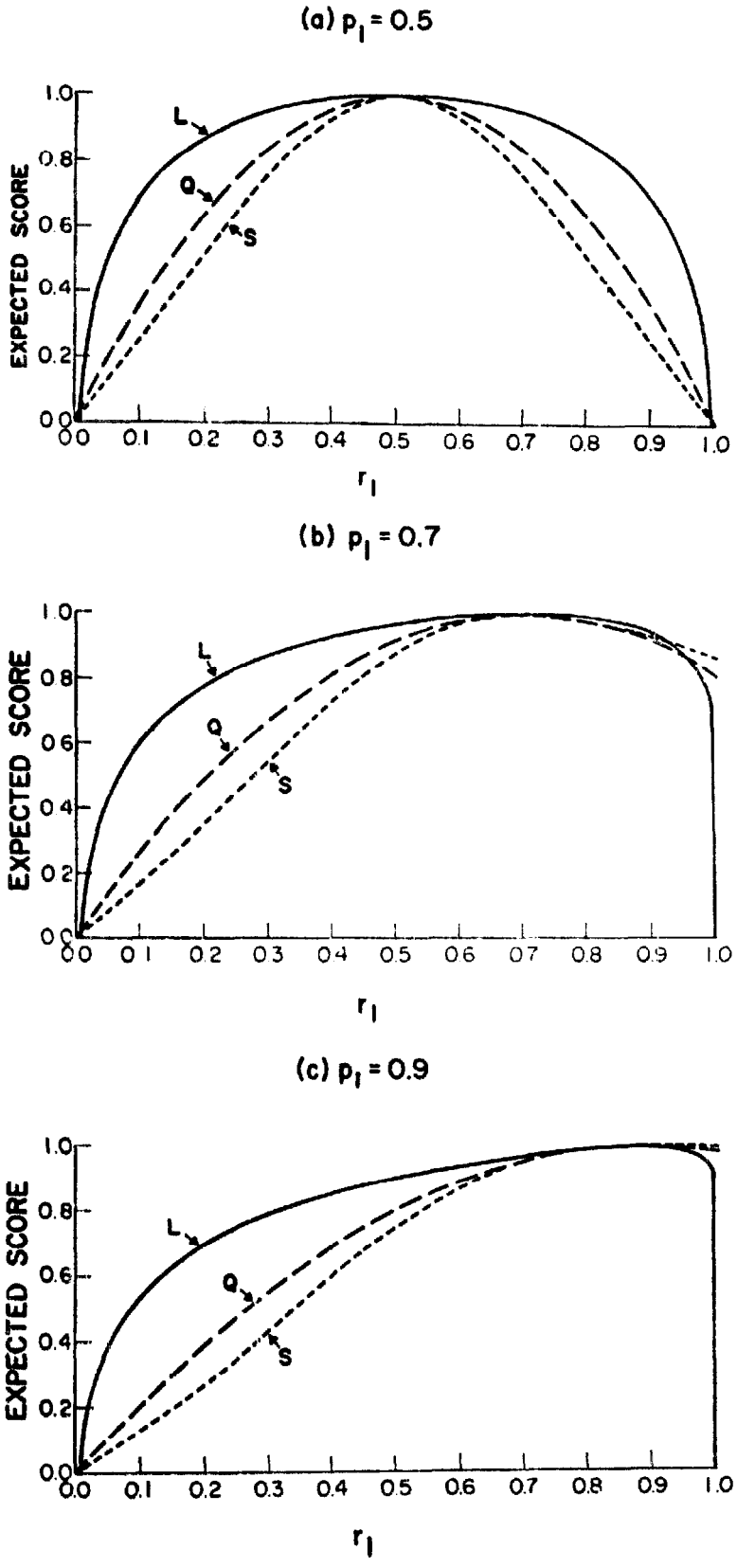

Fig. 1. The expected scores for L, Q and S as a function of r_1 when p_1 equals (a) 0.5, (b) 0.7, and (c) 0.9.

The results for p_1 equal to 0.5, 0.7, and 0.9 are presented in fig. 1.[2] Note that in each case, L seems to be the least sensitive of the three rules to deviations from the optimal forecast $r = p$ (the expected score is 'flatter' for L than it is for Q or S). Also, Q tends to be slightly 'flatter' than S. It appears that for small deviations of r from p in the two-state case, all three of the rules considered are fairly insensitive. If $p_1 = 0.5$, for example, values of r_1 as small as 0.4 or as large as 0.6 result in reductions in the expected scores of approximately 1%, 4%, and 7% for L, Q and S, respectively. Furthermore, as p_1 approaches zero or one, the rules become even less sensitive. If $p_1 = 0.9$, a forecast of $r_1 = 0.6$ only reduces the expected scores by approximately 6%, 11%, and 13% for L, Q and S, respectively.

It is difficult to generalize the above sensitivity results beyond the two-state case. In the general n-state case, the logarithmic rule should still be quite insensitive, since it only depends on r_h, the stated forecast corresponding to the event which actually occurs. The quadratic and spherical rules, however, depend not only on r_h, but also on the distribution of the remaining probability $1 - r_h$ among the remaining $n - 1$ possible outcomes.

For $n = 3$, the quadratic scoring rule was compared with the ranked probability score (RPS).[3] As in the two-state case, the SRs were normalized for a given p to make the maximum and minimum expected scores

---

[2] In order to put the three SRs on a comparable basis, they were normalized for a given p_1 as follows. The maximum value of the expected score, which occurs at $r_1 = p_1$, was set equal to one; the minimum value of the expected score, which occurs at $r_1 = 0$ if $p_1 \geqslant 0.5$ and at $r_1 = 1$ if $p_1 < 0.5$, was set equal to zero. Since L is unbounded below, the minimum value of L was assumed to occur at $r_1 = 0.01$ if $p_1 \geqslant 0.5$ and at $r_1 = 0.99$ if $p_1 < 0.5$. For each SR, this normalizing procedure amounts to a positive linear transformation of the SR. The transformed SRs are strictly proper, since any linear transformation of a strictly proper SR is itself strictly proper.

[3] RPS, unlike the other three rules, is a strictly proper rule which takes into account the ordering of the possible outcomes (provided, of course, that at least ordinal measurement has been attained) (see EPSTEIN, 1969). That is, RPS depends not only on r_h and on the numerical values of the remaining r_i's, but also on the 'closeness' of each potential outcome E_i to the actual outcome E_h. For example, in a football game, if E_1 = win, E_2 = tie, and E_3 = lose, then the forecast $r' = (0.1, 0.4, 0.5)$ receives a better score (using RPS) than $r'' = (0.4, 0.1, 0.5)$ if E_3 occurs, since a tie is 'closer' to a loss than is a win. Under Q, the forecasts r' and r'' would receive the same score. When $n = 2$, RPS is equivalent to Q, since order becomes irrelevant in the two-state case.

equal to one and zero, respectively (see footnote 2). We have examined the relative sensitivity of Q and RPS for a few specific choices of p. The results indicate that the relative sensitivity is quite variable; in some situations, RPS is more sensitive than Q, while in other situations the opposite is true. The relative sensitivity clearly depends upon p. Furthermore, even for a given p, RPS is more sensitive than Q for some values of r and less sensitive for other values of r. In particular, if r_2 is fixed, RPS is least sensitive relative to Q when $r_1 = r_3 = (1 - r_2)/2$ and most sensitive relative to Q when $r_1 = 0$ or $r_3 = 0$. These results are of a preliminary nature, and we intend to examine the relative sensitivity of RPS and Q in more detail. It is evident, however, that neither SR is clearly more sensitive than the other, and that in general the relative sensitivity will depend upon the particular situation.

The sensitivity problem may not be too serious, since the situation may be modified by means of a linear transformation of the SR of concern. As indicated in footnote 2, a positive linear transformation of a strictly proper SR is itself strictly proper. Multiplying the logarithmic rule by a constant greater than one, for instance, will not reduce the insensitivity in *proportional* terms; the proportional reduction in expected score due to deviations will not be changed. However, it will reduce the insensitivity in *absolute* terms, which should encourage more careful assessment. Thus, it is possible to vary the sensitivity of SRs by using appropriate linear transformations. It should be noted, however, that this may increase the potential scores to the point at which the assessor's utility function becomes nonlinear.

There may be psychological considerations in the choice of a particular SR. Even though all of the strictly proper rules should encourage honesty, some may be more likely to encourage honesty in actual practice than others. From a psychological point of view, SRs which can be expressed in relatively simple forms may be preferable to more complicated rules, since the complicated rules are more difficult for the assessor to understand. With a simple rule, it is easier for the assessor to see the relationship between his forecast and his score. Also, the sensitivity question may be related to psychological factors. Certain features of particular SRs, such as the fact that the logarithmic rule is not bounded below, may cause psychological difficulties. At any rate, the choice of a particular SR is not a simple matter, even in the situation in which the assessor's utility function is linear.

3.2. *Nonlinear utility*

If the assessor's utility function is nonlinear, it may not be optimal for him to maximize the expected score. A nonlinear utility function implies that 'risk' considerations may be relevant in the choice of r.

The effects of certain nonlinear utility functions on the quadratic SR, Q, have been considered in some detail by the authors (WINKLER and MURPHY, 1970). If the utility function for positive scores is quadratic, with $U(x) = x^2$ for $x \geq 0$, then the assessor is a 'risk-taker.' The expected utility to the assessor is

$$EU(r) = [E(Q)]^2 + V(Q),$$

where $E(Q)$ and $V(Q)$ are the expectation and variance of the quadratic score. Thus, the expected utility depends not only on the expected Q, but also on the variance of Q. In a two-state situation, the relation of the optimal r_1 to p_1 is illustrated in fig. 2a. Because of the convex utility function, the assessor's stated probability should be closer to the nearest end point, zero or one, than is the actual probability, p_1.

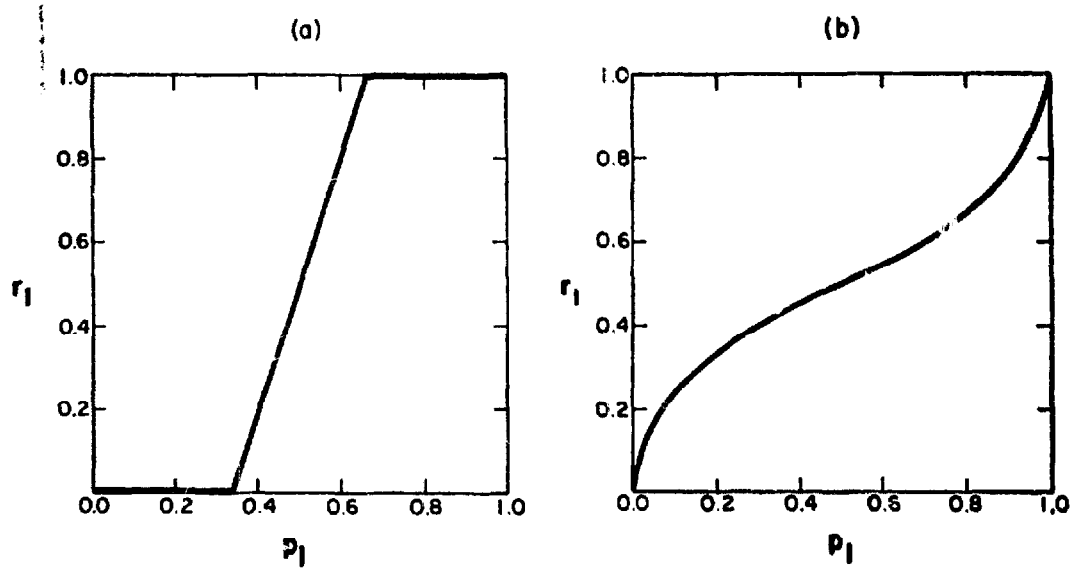

Fig. 2. The optimal r_1 as a function of p_1 in the two-state situation for (a) the 'risk-taker' and (b) the 'risk-avoider'.

If the utility function for positive scores is exponential, with $U(x) = 1 - e^{-x}$ for $x \geq 0$, then the assessor is a 'risk-avoider'. The expected utility is

$$EU(r) = 1 - E(e^{-Q}).$$

In the two-state situation, the relation of the optimal r_1 to p_1 is illustrated

in fig. 2b. Because of the concave utility function, the assessor's stated probability should be closer to 0.5 than is the actual probability, p_1.

What are the implications of nonlinear utility functions for the process of probability assessment (with regard to the utilization of SRs)? If the assessor is able to specify his utility function, then this function can, and should, be incorporated into the assessment process. This is accomplished by defining a new SR, a composite function of the original SR and the (nonlinear) utility function. This composite SR is strictly proper under the given utility function (WINKLER, 1969). Thus, the assessor should not utilize the original SR and 'hedge', but instead should determine the new (composite) SR and maximize the expected score for this rule.

On the other hand, if the assessor's utility function is not known, then the function cannot, of course, be incorporated into the assessment process. Therefore, the assessor's statements may differ from his judgments. For some utility functions the differences might be quite large, while for others they would probably be fairly small (e.g. for approximately linear utility functions the differences should be small). The basic problem, then, is the determination of the assessor's utility function. One approach is to determine the assessor's utility function through the process of interrogation, i.e., by asking the assessor about his preferences. Another approach is to attempt to determine the assessor's utility function through an analysis of his past behavior in similar situations.

4. THE EVALUATION PROCESS

The role of SRs in probability evaluation is to evaluate, i.e., to measure the substantive 'goodness' of, the probabilities. In this section we examine the evaluation problem from the inferential and the decision-theoretic viewpoints and briefly indicate the relevance of certain SRs within the context of particular frameworks.

4.1. *Inferential viewpoint*

From the inferential viewpoint perhaps the most important attribute of the probabilities is their 'validity', i.e., the association between the probability statements and the actual outcomes. Validity has two aspects, which we denote as *primary* validity and *secondary* validity. Primary validity refers to the correspondence between the statement and the relevant observation on an *individual* basis, while secondary validity refers to the correspondence between collections of identical (or similar) statements and the relevant observed relative frequencies on a *collective*

basis. Thus, primary validity relates to the 'accuracy' of the assessor's individual statements; in the two-state case, for example, if $r_1' = 0.6$ and $r_1'' = 0.5$, then r' is more accurate than r'' if E_1 occurs and less accurate if E_2 occurs. Secondary validity, on the other hand, relates to the 'bias' of collections of the assessor's statements (as indicated by the observed relative frequencies; if the assessor uses $r_1 = 0.4$ on a number of occasions, then the bias can be thought of as the difference between 0.4 and the relative frequency of occurrence of E_1 on these occasions). Other attributes of the probabilities have been identified; however, these attributes do not appear to be as important as validity. For a more detailed discussion of such attributes (including validity), we refer to MURPHY and EPSTEIN (1967) and MURPHY (1969a).

The regular simplex, an equilateral triangle in the three-state situation described by DE FINETTI (1962, 1965) and MURPHY (1969a) and illustrated in fig. 3a, provides a natural framework within which to represent the probability statements and the observations and to measure both primary and secondary validity. In this framework, primary validity is related to the (euclidean) distance between the point which represents the statement and the vertex which represents the relevant observation, while secondary validity is related to the distance between the point which represents a collection of identical (or similar) statements and the point which represents the observed relative frequencies. The quadratic SR, which is equivalent to the square of the distance between the point representing r and the vertex, is, then, a 'complete' measure of primary validity in this framework. On the other hand, other strictly proper SRs such as the logarithmic SR and spherical SR are only 'partial' measures of primary validity. SRs which measure secondary validity are not necessarily strictly proper, since secondary validity is defined with reference to *collections* of statements.

The use of a regular simplex and/or the quadratic SR implies that the distances between the states of the variable of concern, i.e., between the vertices of the simplex, are equal. However, if the variable is ordered, this assumption may not be satisfactory. In such a situation we should perhaps use either an irregular, rather than regular, simplex or a SR which takes order into account. As indicated in footnote 3, the ranked probability score (RPS) is a strictly proper SR which takes order into account. RPS and Q have recently been compared in some detail by MURPHY (1970), who has suggested that RPS appears to be a suitable SR for evaluating probability statements for *ordered* variables.

Other frameworks for evaluation from the inferential viewpoint have been proposed. For example, ROBERTS (1965, 1968) has formulated a Bayesian model in which likelihood ratios (or, equivalently, logarithms of likelihood ratios) are used to compare probability statements and thus to evaluate the probabilities themselves. The log likelihood is simply the logarithmic SR and, as a result, within Roberts' framework the logarithmic rule, L, is perhaps to be preferred to other strictly proper SRs (WINKLER, 1969). Clearly, the choice of a particular SR depends upon the framework within which the evaluation is undertaken.

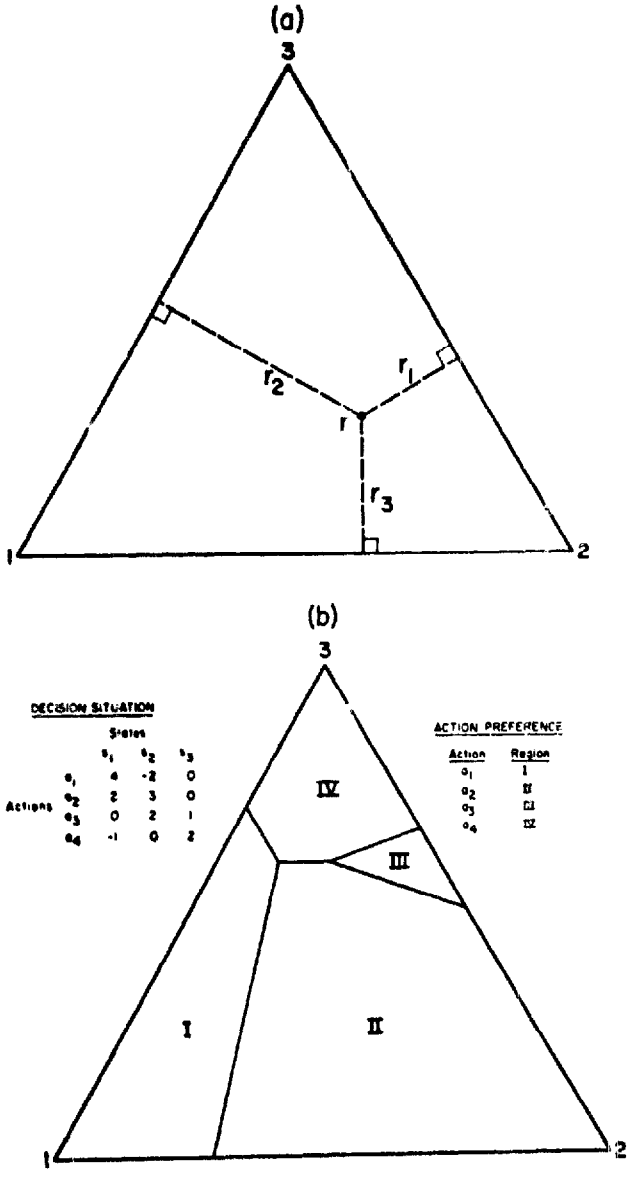

Fig. 3. (a) The regular simplex in the three-state situation. The point $r=(r_1, r_2, r_3)$ represents a particular probability statement. (b) The representation of a particular decision situation within the framework of the regular simplex.

4.2. *Decision-theoretic viewpoint*

In this subsection we describe the general framework within which we consider evaluation from the decision-theoretic viewpoint, and then we indicate briefly the nature of certain results and problems. We assume that the assessor is an expert in a substantive area, e.g., in meteorology, who formulates probability statements (forecasts) and provides such statements to decision makers (DMs) who use the statements in their decision-making problems.

In a decision situation, the consequence, or payoff, to the DM depends on the action which he takes and on the event E_h which actually occurs. In turn, the action which he takes depends on the assessor's stated probabilities. Therefore, the consequence, and hence the utility of the consequence, depends (indirectly) on r. Thus, in this framework, if the DM selects action i and E_h occurs, then the DMs utility is $U_h(r) = u_{ih}$, which is a natural SR for the assessor's statements. It is of interest to note that if we assume that the assessor's utility function is linearly related to the DM's utility function, an assumption which may be reasonable in many situations, then $U_h(r)$ is a proper (but not necessarily strictly proper) SR (MURPHY, 1969a; RAIFFA, 1969). Using $U_h(r)$ as a SR is consistent with a suggestion of SAVAGE (personal communication, 1970) that a DM should reward the assessor with a share of the decision-making problem.

A DM's knowledge of his utilities may be incomplete, in which case he may express his knowledge in terms of probabilities. Such probabilities can be interpreted in several different ways: (1) in a 'one-to-one' situation, in which the assessor's probability statement is used by a single DM, the probabilities express the DM's uncertainty concerning his utilities; (2) in a 'one-to-many' situation, in which the assessor's probability statement is used by several different DMs, the probabilities describe the distribution of the DMs' utilities for money over the different decision problems. When knowledge of the utilities is expressed in probabilistic terms, the *expected* utility, rather than the utility, can be used as a SR (where the expectation is with regard to the distribution of utilities).

The DM's problem can also be represented within the framework of the regular simplex (a particular decision situation is depicted within this framework in fig. 3b). The simplex is divided into regions which correspond to the DM's actions; i.e., if the assessor's statement falls in a particular region the DM selects the action which corresponds to that region.

We briefly describe certain results within this general framework in

the two-state and n-state situations. In the two-state situation the simplex reduces to the unit line segment. The score assigned by any strictly proper SR is a strictly decreasing function of the distance from the end-point of the unit interval which corresponds to the actual event; the utility is a decreasing function of this distance. For example, if $r_1' = 0.8$, $r_1'' = 0.6$, and E_1 occurs, then any strictly proper SR assigns a higher score to r'. The utility resulting from r' is greater than or equal to the utility resulting from r''. As a result, if $r_h' > r_h''$, then $U_h(r') \geqslant U_h(r'')$ (MURPHY, 1969a).

In the cost-loss ratio decision situation (MURPHY, 1966, 1969a, b), a two-action, two-state situation in which the DM's utility matrix contains only one unknown parameter, the cost-loss ratio (in terms of utility), the following statement holds: if the cost-loss ratio is assumed to possess a uniform distribution, then the expected utility SR is a linear function of the quadratic SR. If beta distributions with integer parameter values are considered, the resulting expected utility SRs are polynomials. Since $U_h(r)$ is a proper SR, these expected utility SRs are strictly proper SRs (MURPHY, 1969c).

Recall that, within the geometric framework, (primary) validity is a strictly decreasing function of distance. Utility, on the other hand, is a decreasing function of *directed* distance. That is, the utility of the assessor's statements cannot increase (with distance) along a directed line segment through a vertex with reference to the state corresponding to that vertex.

The RPS was formulated in the context of a specific n-action, n-state decision situation in which the single parameter ('utility') of concern was assumed to possess a uniform (probability) distribution. Although such expected utility SRs are of necessity associated with specific decision situations, they may, like the RPS, be of some general interest. Unfortunately, decision situations, in general, lead to very complicated expected utility SRs. The complexity of the n-state situation is such that general results are difficult to obtain.

REFERENCES

DE FINETTI, B., 1962. Does it make sense to speak of 'Good Probability Appraisers'? In: I. J. Good (ed.), The scientist speculates – an anthology of partly-baked ideas. New York: Basic Books, 357–364.

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

DE FINETTI, B., and L. J. SAVAGE, 1969. The elicitation of personal probabilities and expectations. Unpublished manuscript, University of Rome and Yale University.

EPSTEIN, E. S., 1969. A scoring system for probability forecasts of ranked categories. J. appl. Meteorol. **8**, 985–987.

FISHBURN, P. C., 1964. Decision and value theory. New York: Wiley.

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. Meteorol. **5**, 534–537.

———, 1969a. The evaluation of probabilistic predictions in meteorology. Unpublished Ph.D. dissertation, University of Michigan, Ann Arbor.

———, 1969b. 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. Meteorol. **8**, 863–873.

———, 1969c. On expected-utility measures in cost-loss ratio decision situations. J. appl. Meteorol. **8**, 989–991

———, 1970. The ranked probability score and the probability score: a comparison. Monthly Weather Review **98**, in press.

——— and E. S. EPSTEIN, 1967. Verification of probabilistic predictions: a brief review. J. appl. Meteorol. **6**, 748–755.

RAIFFA, H., 1969. Assessment of probabilities. Unpublished manuscript, Harvard University.

ROBERTS, H. V., 1965. Probabilistic prediction. J. Amer. statist. Ass. **60**, 50–62.

———, 1968. On the meaning of the probability of rain. Proceedings of the First Conference on Statistical Meteorology. Boston: American Meteorological Society, 133–141.

SAVAGE, L. J., 1954. The foundations of statistics. New York: Wiley.

STAËL VON HOLSTEIN, C.-A. S., 1970. Some problems in the practical application of Bayesian decision theory. In: W. Goldberg (ed), Behavioral approaches to modern management. Gothenburg: The Graduate School of Economics and Business Administration.

WINKLER, R. L., 1967. The quantification of judgment: some methodological suggestions. J. Amer. statist. Ass. **62**, 1105–1120.

———, 1969. Scoring rules and the evaluation of probability assessors. J. Amer. statist. Ass. **64**, 1073–1078.

——— and A. H. MURPHY, 1968. 'Good' probability assessors. J. appl. Meteorol. **7**, 751–758.

——— and ———, 1970. Nonlinear utility and the probability score. J. appl. Meteorol. **9**, 143–148.