ORGANIZATIONAL BEHAVIOR AND HUMAN PERFORMANCE 9, 307–317 (1973)

Psychological Effects of Proper Scoring Rules[1]

FLOYD A. JENSEN AND CAMERON R. PETERSON[2]
University of Michigan

Proper scoring rules (PSRs) have been derived to elicit good probability assessments. Because there are so many different kinds of PSRs, this experiment was designed to determine if any particular characteristics contribute to effectiveness. Subjects observed poker chips in jars and bet on the color of the chip to be sampled. On different trials, lists of bets were generated by different PSRs. The type (log, quadratic, or spherical) of PSR used appeared to have essentially no effect on the probability inferred from the bet selected. However, the inferred probability became less extreme with increased steepness in the functions relating score to assessed probability. Also, various suboptimal strategies seemed to be employed when the rule contained both positive and negative scores, so all possible scores should probably be either positive or negative but not both.

There are many uncertain events in the world, and often people evaluate the degree of uncertainty in their own minds about an event by stating a probability. The practice of attaching probabilities to uncertain events is well known in the field of meteorology, where we are told, for example, that the "chance of rain tomorrow is 30%," but probabilistic assessments are also used in such diverse fields as business, medicine, intelligence, and psychology, not to mention the world of gambling.

Many investigators have broached the problem of assessing a person's subjective probability distribution (SPD) over an event space so that it is not hedged one way or another (e.g., Savage, 1971; Murphy & Epstein, 1967; Winkler, 1967; Winkler & Murphy, 1968a). For example, a weather forecaster may believe, after examining all available data, that the probability of rain tomorrow is .50, yet if he is biased in favor of predicting rain (perhaps because farmers need it), he may forecast .70 as the probability of rain.

[1] The research reported here was undertaken in the Engineering Psychology Laboratory, Institute of Science and Technology, University of Michigan. The research was supported by the Wood Kalb Foundation.

[2] Requests for reprints should be sent to Cameron R. Peterson, Engineering Psychology Laboratory, Highway Safety Research Institute, Institute of Science and Technology, University of Michigan, Ann Arbor, MI 48105.

Copyright © 1973 by Academic Press, Inc.
All rights of reproduction in any form reserved.

A class of payoff functions known as proper scoring rules has been considered as one approach to encourage probability assessors to state their true beliefs. In general a scoring rule is any algorithm that assigns a payoff for a probability assessment, where the payoff depends only on the assessor's stated probability distribution *and* the event that actually occurs. A *proper* scoring rule (PSR) is a payoff function whereby a person can maximize his subjectively expected score only by stating his true beliefs. Letting $p = (p_1, p_2, \ldots, p_n)$ represent a person's SPD over n mutually exclusive and exhaustive hypotheses and $r = (r_1, r_2, \ldots, r_n)$ his stated (assessed) distribution, under a PSR he can maximize his subjectively expected score by setting $r = p$.

The three most common "types" of PSRs are the logarithmic, the quadratic, and the spherical. Letting $S_k(r)$ denote the score which obtains if the kth event occurs, the three rules may be stated mathematically as follows:

$$\text{logarithmic: } S_k(r) = \log r_k$$

$$\text{quadratic: } S_k(r) = 2r_k - \sum_{i=1}^{n} r_i^2$$

$$\text{spherical: } S_k(r) = r_k \Big/ \left(\sum_{i=1}^{n} r_i^2 \right)^{1/2}$$

It has been demonstrated elsewhere (e.g., Winkler & Murphy, 1968b) that the above three rules are indeed proper.

Any PSR remains proper through a linear transformation, provided the multiplicative constant is positive (Toda, 1963). Thus, it is possible to generate an infinite number of different PSRs of any one type, and these three are by no means the only types.

Which of the infinite number of PSRs should one use? On purely theoretical grounds it should make no difference. The rationale behind the use of PSRs is that a person should evaluate his subjectively expected value (SEV) function for the different probabilities he can assess, and then state the probability which maximizes that function. Therefore every proper scoring rule should have the same effect; the person should state $r = p$ in order to maximize SEV.

However, different scoring rules may have different psychological effects. For example, the "sensitivity" or penalty for hedging assigned by the logarithmic, quadratic, and spherical PSR has been investigated by Murphy and Winkler (1970b). It turns out that the SEV functions tend to be comparatively flat in the regions of their maxima, so that a person is penalized relatively little in terms of SEV for deviating sub-

stantially from stating his true SPD. Thus, it would seem that stated probabilities could be pushed around by a variety of factors, even when a PSR is used. The problem of the "flat maximum" can be accentuated or lessened through multiplicative transformations of any of the PSRs. Such a transformation merely changes the range of possible scores, but such changes can make the SEV function almost perfectly flat or relatively steep, depending on the multiplicative constant.

Besides multiplication by a constant, one can also perform a linear transformation on any of the PSRs by adding or subtracting a constant from every score. For example, it could be arranged so that all scores would be positive outcomes (which requires truncation with the log rule), negative outcomes, or a mixture of the two. It may be that both additive and multiplicative manipulations have psychological consequences on probability statements.

It may also be that the characteristic shape of each type of PSR has an effect on assessments. For example, the logarithmic rule is very steep close to $r_k = 0$ and flattens out as r_k increases, becoming almost linear between $r_k = .5$ and $r_k = 1.0$. The quadratic rule, as the name implies, always has a parabolic shape, while the spherical rule has a very slight S shape, being almost horizontal near $r_k = 1.0$. Arguments have been made on both theoretical (Raiffa, 1969; Winkler, 1969) and empirical (Phillips & Edwards, 1966) grounds that the logarithmic scoring rule is superior to the others. Meteorologists, on the other hand, seem to favor the quadratic rule, at least in practice. Murphy and Winkler (1970b) have considered in detail the relative merits of different PSRs.

Because of the increasing use of PSRs, it is becoming important to investigate the differential psychological effects of using different scoring rules. This is the problem addressed by the present experiment. Specifically, what is the psychological effect of varying the type or transformation of a PSR?

METHOD

Experimental Design

In most practical situations the assessor's subjective probability for an event is not known, so it is difficult to know whether his stated probability is his "best" estimate. In the present experiment, however, there was not as much interest in what the subjective probabilities were as in evaluating how *S*s would react to different kinds of PSRs. Therefore, the correct probabilities were displayed as transparently as possible and *S*s were given extensive experience in dealing with these probabilities so that it was reasonable to assume that their subjective probabilities would

conform to the correct or "public" probabilities. The *S*s were not asked to state r but rather they were asked to focus on the payoffs generated by the PSR and select the set of payoffs they considered most desirable. This was accomplished by having the *S*s select one bet from a list of bets for each PSR. Values of r were inferred from their choices of bets. Differences in inferred rs resulting from different scoring rules provided a means of measuring biases attributable to different PSRs. In this experiment we consider only the case of two possible events. The *S* can be thought of as making only a single probability assessment, e.g., the probability of the favored hypothesis; we shall call that assessment r_j. Additivity of probabilities determines the other value. Two binomial populations were used, one with proportions .60 and .40, the other with proportions .85 and .15, in order to test PSRs under two levels of objective probabilities.

For each of the two populations each of the three types of PSRs was used to generate two lists of 50 bets each in which the inferred probabilities for r_j ranged from .05 to .98. Scale constants for the functions generating each list were determined by the following set of criteria: first, the EV for estimating the probabilities correctly was set at 50 points. Thus, every EV function passed through the same maximum point for each population. Second, for each PSR two conditions were created: under the Flat condition, an estimate of r_j of .05 would yield a score of -20 points if the other event occurred; in the Peaked condition an r_j of .05 would yield a score of -160 points if the other event occurred.

The above constraints had two major effects. First, they forced different types (log, quadratic, and spherical) of PSRs to be similar in steepness. Second, these scale transforms forced two versions of each type of PSR to be strikingly different from each other in steepness. As the labels imply, under the Flat condition the EV function was much flatter than under the Peaked condition.

As a control condition, to insure that *S*s were paying attention to the task, linear scoring rules were used to generate lists of bets for each of the conditions described above. The optimal strategy under a linear scoring rule is to estimate $r_j = 1.0$, because that will maximize SEV. Previous research (e.g., Phillips & Edwards, 1966) has demonstrated that linear scoring rules elicit more extreme r_j than do PSRs. Scale constants for the linear rules were determined by the same criteria as for the PSRs.

The resulting scoring rules are displayed in Fig. 1. In addition to these scoring rules, two additional PSRs were created by either adding or subtracting 50 points from every bet in the "flat" logarithmic list for the 60–40 population. This preserved the properness of the rules while per-

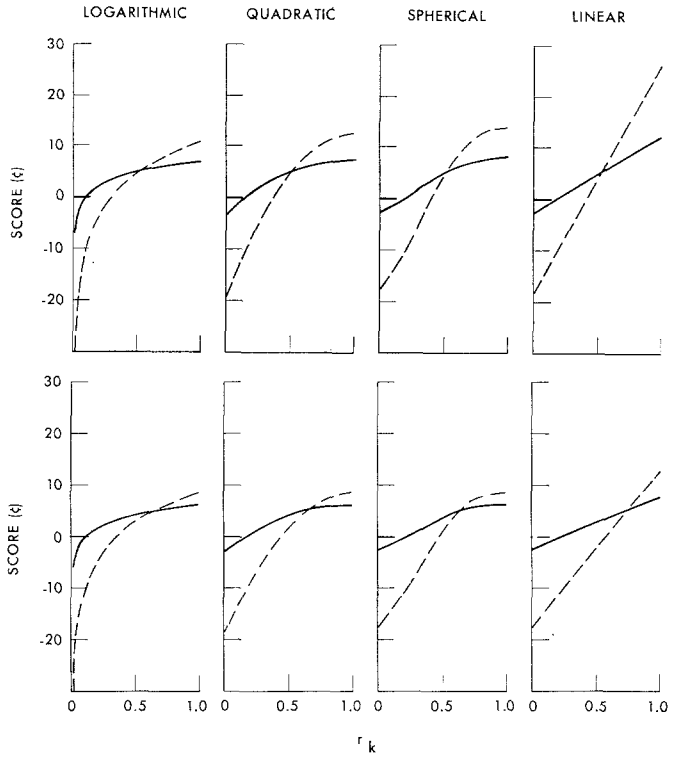

FIG. 1. Scores under various scoring rules as a function of r_k, the estimated probability for the event that occurs. (Solid lines refer to the Flat condition; dashed lines refer to the Peaked condition. The upper graphs are for the 60–40 population; the lower graphs are for the 85–15 population.)

mitting a comparison of three PSRs which were identical in form, differing only in overall EV.

Procedure

The first part of the experimental session was played for practice, and in the second part the bets were played for real, i.e., for money. Subjects were run individually and received no specific training in the use of any strategy, in fact, they were encouraged to experiment with different strategies during practice. At the beginning of the experiment each *S* was shown a large, transparent glass jar filled with poker chips of two colors, representing either the 60–40 or the 85–15 population (the 60–40 population consisted of 60% blue chips and 40% red chips and the 85–15 population consisted of 85% white chips and 15% red chips). The *S* did

not know the proportions beforehand. After careful visual examination of one population, *S* was asked to estimate the proportions of that population, following which he was told what the true proportions were. Those proportions were thereafter displayed on cards in percentage form to remind *S*s of the correct proportions.

The *E* then emptied that jar into a bookbag, and practice commenced. The lists of bets for that population were arranged in random order in a booklet, one list to a page. From each list *S* chose the bet he most wanted to play and indicated his choice verbally to *E*. After each choice the bet was played; play consisting of having *S* draw a poker chip from the bookbag. The color of the chip determined *S*'s winnings for that bet. This process was repeated until all bets for that population had been played. Then the *S* followed the same procedure for the other population. Half of the *S*s started with the 60–40 population and half started with the 85–15 population. The practice part of the experiment was intended to give *S*s familiarity with the procedure, with the proportions involved, and with the sample outcomes of the bets they chose. The *E* kept track of winnings for the practice session and at the conclusion of practice told *S* the total, the exchange rate from points to money (1 point = .1 cent), and how much *S* would have won had the practice bets been played for real. Then the real session started, proceeding in the same manner as the practice session with two exceptions: For each *S* the order of presentation of lists for each population was randomized; and *S* made his choices by marking his preferred bet in the booklet. At the conclusion of the experiment *S* was asked to respond in writing to two questions: (a) "What did you think the experiment was about?" and (b) "If you used any kind of strategy in choosing your bets, what was it?"

Subjects

Twenty-two men students from The University of Michigan served as *S*s. A *S*'s pay was determined by the outcomes of the 18 bets played during the 1-hr experiment; possible earnings ranged from −$1.05 to $2.54. Actual earnings ranged from $.94 to $1.90 with a mean of $1.54.

RESULTS

Subjects' verbal estimates of the proportions for the two populations indicated that they had good intuitive understanding of the proportions involved. For the 60–40 population the mean estimate of the proportion of the predominant color was .61 (SD = .07). For the 85–15 population the mean estimate of the proportion of the predominant color was .83 (SD = .08).

Figure 2 displays the results of individual *S*s in conjunction with ex-

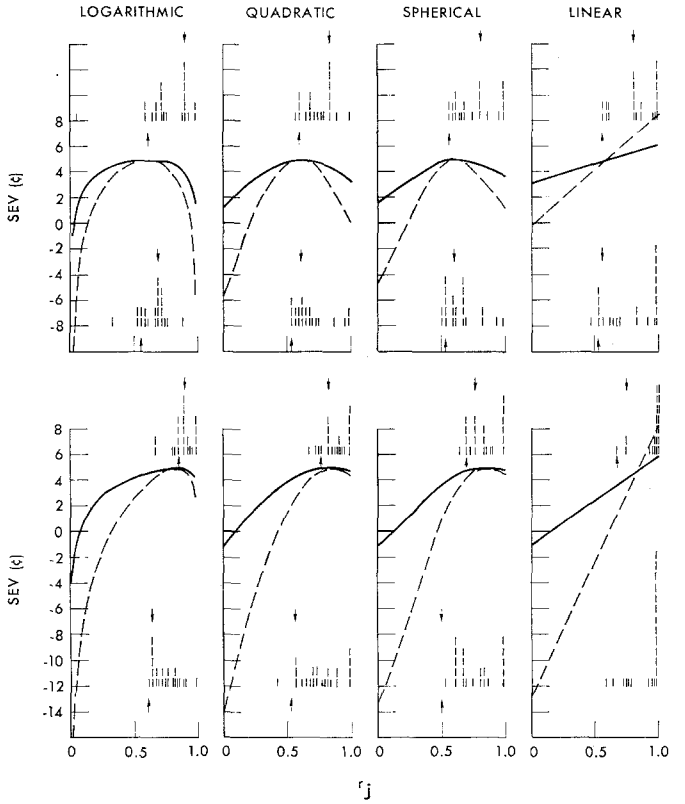

FIG. 2. Distribution of *Ss'* inferred probability estimates for the more likely hypothesis (r_j), plotted on SEV functions derived from the scoring rules in Fig. 1. (Points above the curve are for the Flat conditions; points below the curve are for the Peaked condition.)

pected value functions corresponding to the scoring rules displayed in Fig. 1. The expected value functions are related to the scoring rule functions by the formula $EV = p_a S_a(r) + p_b S_b(r)$, where a and b refer to the two possible events. Observe that manipulation of the multiplicative constant produced substantial differences in steepness of the EV functions between the flat and the peaked conditions. The *Ss'* inferred probability assessments, r_j, were calculated by using the respective scoring rules. Those inferred values are plotted to show the distribution of responses across possible values of r_j. Points above the curves refer to the flat condition; points below refer to the peaked condition.

It is obvious that *Ss'* responses to linear scoring rules were substantially more extreme than their responses to PSRs; this difference is statistically significant at the .05 level as measured by t tests. This result

reproduces previous findings (e.g., Phillips & Edwards, 1966) and indicates that *Ss* were attending to the task, i.e., that PSRs did have some influence on their responses. A summary of mean inferred probability assessments for most of the scoring rules tested is given in Table 1.

It is clear that *Ss* responded more extremely to the 85–15 populations. However, for the 60–40 population both the mean flat estimate for PSRs (.76) and the mean peaked estimate (.65) overshot the correct probability, whereas for the 85–15 population the mean flat estimate (.86) was very close to optimal, but the mean peaked estimate (.75) undershot the correct probability considerably.

With respect to differences between types of PSRs, Figure 2 indicates that under the sets of scale constraints considered, there were essentially no differences between mean inferred assessments between any of the types of PSRs. This conclusion was again confirmed by statistical tests (two-tailed t tests at the .05 level were used throughout these results); of 12 comparisons between means, only two showed significant differences, those two occurring in the "flat" condition for the 85–15 population, where the spherical rule yielded probability assessments significantly lower than the quadratic or logarithmic rules. The differences among characteristic shapes of different types of PSRs do not appear to appreciably influence *Ss'* probability assessments.

With respect to the effect of manipulating the steepness of the PSRs through multiplicative transformations, it was expected that sharpening the SEV function by making the PSR steeper would cause a reduction in the variability of *Ss'* responses, since deviating from the optimal assessment would be much more costly in terms of SEV under a peaked rule. However, as can be observed in Figure 2, no such systematic reduction of variance occurred. Surprisingly, varying "steepness" appeared to have an effect on central tendency, but not on variance. The combined

TABLE 1

MEAN INFERRED PROBABILITY ESTIMATES FOR THE MORE LIKELY HYPOTHESIS

Scoring rule	60–40 Population			85–15 Population		
	Flat		Peaked	Flat		Peaked
Logarithmic	.77	(.05)[a]	.65	.88	(.02)	.75
Quadratic	.76	(.05)	.65	.88	(.04)	.75
Spherical	.75	(.03)	.65	.83	(.03)	.76
Linear	.81	(.03)	.78	.94	(.04)	.90

[a] Numbers in parentheses are the standard errors of the difference between flat and peaked conditions.

mean inferred probability assessment for flat PSRs was significantly greater than the mean assessment for peaked PSRs, both for the 60–40 and for the 85–15 populations. In the six individual comparisons between flat and peaked forms of the same PSR, all but one indicated significant differences. Thus, it is clear that manipulating the multiplicative constant in a PSR does affect *S*s' inferred probabilities even though not in the way expected.

The additive constant also had an effect on *S*s' responses. There was no significant difference between mean assessments for the flat logarithmic rule (.77) and the same rule with 50 points added to each score (.73), but the mean assessment for the same rule with 50 points subtracted from each score was significantly lower than for the other two rules. In fact, even though that PSR offered the worst set of bets in terms of expected value, it produced a mean probability assessment closest to the correct proportions of any of the scoring rules (.61).

DISCUSSION

Which scoring rule should one use? If the probability assessor maximizes SEV, it should make no difference. However, as has been demonstrated here, people apparently do not carefully evaluate and maximize their SEV function. Other factors influence assessments.

It appears to be relatively unimportant which *type* of PSR one uses, at least for the populations tested (with extreme probabilities such as between .99 and 1.0 one might well expect the logarithmic rule to produce assessments different from the spherical and quadratic rules, since the SEV function for the logarithmic rule falls off to $-\infty$ near $r_k = 1.0$ and $r_k = 0$). The characteristic shape of each type of PSR does not seem to influence probability assessments to any large extent.

However, multiplicative transformations of PSRs do have large effects on mean probability assessments, although, contrary to expectations, they had no reliable effect on the variance of probability assessments. Making the PSR function steeper seems to make probability assessments more cautious, to move them nearer to .50. There are several alternative explanations of this phenomenon. One is that the penalties for being "wrong" with extreme responses are much more severe with the steeper PSRs. A value of r_k near 1.0 for the favored event implies a value near 0 for the other event. Note in Figure 1 how low all the steep functions go as r_k approaches 0. The *S*s may have been hesitant about taking chances with such large losses.

However, there are other possible explanations for the effect of the multiplicative transform. Both from their answers to the question "What

was your strategy?" and from their choices, one can discern that *Ss* sometimes looked for the bet whose two possible outcomes formed a ratio close to that of the population proportions. Such a strategy is not in general optimal, but it may be intuitively compelling.

Nevertheless, that strategy was apparently not employed as often as another nonoptimal scheme. The *Ss* often chose the bet which had the smallest positive value attached to the less likely hypothesis, such that if that hypothesis was not correct, then they would still lose nothing. The *Ss* often seemed to express a reluctance to take a bet where there was a possibility of a loss, so a discontinuity appeared where values changed from positive to negative. This last strategy may provide at least a partial explanation for the differences between means of the Flat vs Peaked conditions. In the Flat condition the bet in each list one would choose if using this strategy always corresponded to a higher inferred probability than the bet from the Peaked list which that strategy would lead one to choose.

To get an idea of the extent to which the last two strategies described above were used, observe the distribution of *Ss'* probability assessments in Fig. 2. The arrow pointing up indicates the inferred probability for that bet for which the ratio of values best approximated the ratio of the population proportions. The arrow pointing down indicates the inferred probability for that bet which offered the smallest positive value for the less likely hypothesis (given that the value for the more likely hypothesis was positive). Where arrows do not appear there was no bet in that list which fitted one or the other of the above conditions. It is interesting to note that the list that produced the best assessments did not contain any bets one could choose by the above simple strategies. This leads one to wonder whether people would be more willing to maximize SEV if alternative strategies were less available or obvious.

One way of eliminating difficulties encountered in changing from positive to negative scores is through additive transformations, so that all possible scores are either positive (recall that the log rule requires truncation) or negative, but not a mixture. In this context it is interesting to note that one industry which regularly makes use of PSRs, meteorology, uses a form of the quadratic score where all outcomes are negative (the "Brier" score) and the optimal strategy is to minimize losses.

In conclusion, these results have rather complex implications about which scoring rule should be used for the purpose of motivating better probability assessments. The form of the PSR is essentially irrelevant, at least over the large range of the probability scale. It is probably a good idea to restrict to all positive values or all negative values in order to counter the use of some nonoptimal strategies. Finally, the use of

a more peaked PSR seems to lead to more cautious assessments, but it is difficult to know in advance whether that is a desirable or undesirable quality.

REFERENCES

MURPHY, A. H. A note on proper and strictly proper scoring rules. Unpublished manuscript, University of Michigan, 1969.

MURPHY, A. H. The ranked probability score and the probability score: A comparison. *Monthly Weather Review,* 1970, **98,** 917–924.

MURPHY, A. H., & EPSTEIN, E. S. A note on probability forecasts and "hedging." *Journal of Applied Meteorology,* 1967, **6,** 1002 1004.

MURPHY, A. H., & WINKLER, R. L. Scoring rules: A bibliography. Unpublished manuscript, University of Michigan and Indiana University, 1970. (a)

MURPHY, A. H., & WINKLER, R. L. Scoring rules in probability assessment and evaluation. *Acta Psychologica,* 1970, **34,** 273–286. (b)

PHILLIPS, L., & EDWARDS, W. Conservatism in a simple probability inference task. *Journal of Experimental Psychology,* 1966, **72,** 346–354.

RAIFFA, H. Assessment of probabilities. Unpublished manuscript, Harvard University, 1969. Cited by Stael von Holstein, C. A. S. Measurement of subjective probability. *Acta Psychologica,* 1970, **34,** 146–159.

SAVAGE, L. J. Elicitation of personal probabilities. *Journal of American Statistical Association,* 1971, **66,** 783–801.

STAEL VON HOLSTEIN, C. A. S. A family of strictly proper scoring rules which are sensitive to distance. *Journal of Applied Meteorology,* 1970, **9,** 360–364.

TODA, M. Measurement of subjective probability distribution. Institute for Research, Division of Mathematical Psychology, State College, Pennsylvania State University, 1963.

WINKLER, R. L. The quantification of judgment: Some methodological suggestions. *Journal of American Statistical Association,* 1967, **62,** 1105–1120.

WINKLER, R. L. Scoring rules and the evaluation of probability assessors. *Journal of American Statistical Association,* 1969, **64,** 1073–1078.

WINKLER, R. L., & MURPHY, A. H. Evaluation of subjective precipitation probability forecasts. In *Proceedings of the first national conference on statistical meteorology,* Boston: American Meteorological Society, 1968. Pp. 148–157. (a)

WINKLER, R. L., & MURPHY, A. H. "Good" probability assessors. *Journal of Applied Meteorology,* 1968, **7,** 751–758. (b)

RECEIVED: May 20, 1972