ORGANIZATIONAL BEHAVIOR AND HUMAN PERFORMANCE 6, 304–315 (1971)

The Effect of Learning on the Assessment of Subjective Probability Distributions[1,2]

CARL-AXEL S. STAËL VON HOLSTEIN
University of Stockholm and The Economic Research Institute at the Stockholm School of Economics

Subjects were asked to assess subjective probability distributions for unknown parameters of a Bernoulli process. The process was generated by random devices such as oddly shaped dice. One group of subjects received feedback between the sessions in the form of the true values of the parameters. The members of this group improved their performance more than the second group. They also acquired more confidence in their assessments.

1. INTRODUCTION

In almost all applications of Bayesian methods in the literature, it is assumed that there exists a prior distribution and that this distribution is known. The research is mainly directed towards questions that arise when prior distributions are incorporated into the analysis of a problem. An important field of research treats the mathematical problem of transforming the prior distribution into a posterior distribution by additional knowledge in the form of sample information (Raiffa & Schlaifer, 1961; Edwards, Lindman, & Savage, 1963). Psychologists are interested in the way people process such information, i.e., how well their posterior distributions agree with the ones obtained by a formal Bayesian model (Phillips & Edwards, 1966; Edwards, Phillips, Hays, & Goodman, 1968).

The problem of how to assess these prior distributions, or subjective probability distributions (SPDs), is seldom discussed. They are subjective in that they reflect the assessor's personal comprehension of certain events and measure his degrees of belief in these events. The choice of assessment technique is not obvious, and experiments have shown that subjects react in different ways when confronted with different techniques. It may also be that a technique which is satisfactory

[1] This study was supported by grants from The University of Stockholm and from The Bank of Sweden Tercentenary Fund.

[2] This is a revised version of a paper presented at the European Meeting on statistics, econometrics, and management science sponsored by IMS, TIMS, ES, and IASPS, Amsterdam, 2–7 September 1968.

in one situation might not work well in another situation. There is a need not only for different techniques for the assessment of SPDs, but also for empirical evidence of how well they work in practice. Empirical research may be expected to lead to improved and more reliable techniques. The only extensive study of the problem of assessing SPDs was carried out by Winkler (1967a), who discusses several assessment techniques and also presents empirical results.

It is conceivable that a person may be trained to make assessments that are better in some sense. We shall not go into the problem of defining what should be meant by a good assessment but refer instead to discussions by de Finetti (1962), Winkler (1967b), and Winkler and Murphy (1968). It suffices here to say that improved assessments may be due to increased knowledge of the problem area or to a better understanding of the assessment technique(s). Training is of great importance when the assessments are to be utilized in a practical decision situation. The most suitable form of training probably differs from one application to another. Results from empirical research might be helpful when designing the training for a particular case.

The main purpose of the present experiment was to study the effect of one simple kind of training. The experiment was divided into three sessions, and training was provided in the form of feedback of results between the sessions. The goodness of an assessment was measured by a scoring method based on the assessed SPD and the true probability. An improvement in the scores as the experiment progressed would indicate a positive effect of the feedback. This effect is, of course, related to the assessment technique used and could be different with some other technique.

The study is restricted to the assessment of SPDs for unknown probabilities or parameters of a Bernoulli process. The participants were shown various random devices such as two-colored irregular dice made of homogeneous material. They were asked to estimate the probability that the device would produce a certain event, e.g., that after a throw, the die would come to rest on a red surface. It was assumed that the exposure of the device would give the participants some information about the behavior of the device which would lead to a SPD that was neither diffuse nor tight.

2. THEORETICAL BACKGROUND

Form of Assessed Distributions

Any distribution defined on the interval [0, 1] could serve as a SPD. However, it may be convenient for computational reasons, as in the present study, to approximate the SPD by a member of some family of distributions. It could also be argued that an assessor cannot specify every detail of his SPD and that he probably cannot distinguish between two similar distributions (i.e., he may be willing to act in accordance with the implications of either of these two distributions). A suitable choice in the present situation is the family of β-distributions with density functions,

$$f(p) \propto p^{a-1}(1-p)^{b-1}, \qquad 0 \leq p \leq 1, \qquad a > 0, \qquad b > 0.$$

They have a mathematically simple form; yet, the two procedures should provide enough flexibility to give a good approximation of most reasonable distributions. Some exceptions were found in the experimental results, and these will be discussed in Section 4.

There are several possible ways of eliciting information from subjects in order to determine their SPDs (Pratt, Raiffa, & Schlaifer, 1965; Winkler, 1967a). Because of the experimental conditions with several people participating in the same session, it was thought that matters would be simplified by having a technique which was both easily administered and based on well-known statistical concepts. It was, therefore, decided that the subjects should state the median and the two quartiles of their SPDs. This provides one condition too many for the fitting of a β-distribution and the parameters a and b were obtained by the following approximate method, proposed by Pratt *et al.* (1965, Section 11.5).

Let m be the median and q_i the i'th quartile of a β-distribution, and let

$$a_i = \{[m(1-q_i)]^{1/2} - [q_i(1-m)]^{1/2}\}^2, \qquad i = 1, 3.$$

The parameters of the β-distribution are then given approximately by

$$a = km + \frac{1}{3} \quad \text{and} \quad b = k(1-m) + \frac{1}{3}, \qquad \text{where } k = .056\left[\frac{1}{a_1} + \frac{1}{a_3}\right].$$

Evaluation of Assessments

Each assessment was measured by a score which was based on the assessed SPD (in the form of the fitted β-distribution) and the true probability. The scores from one session were used as feedback in the following session. The scoring rule used is a variation of the quadratic scoring rule, adapted to continuous probability distributions. (For an extensive treatment of the choice of suitable scoring methods see Winkler, 1967b). Let p denote the true probability. The interval [0, 1] is then divided into seven subintervals, the dividing points being p — 0.06, p —

0.03, $p - 0.01$, $p + 0.01$, $p + 0.03$, $p + 0.06$.[3] Let r_k denote the mass of the β-distribution in the k'th interval. r_k is then an approximation of the subject's assessed probability for the k'th interval. The score S was set to $0.5(S_1 + 1)$ where $S_1 = 2r_4 - \Sigma r_k^2$. The highest possible score is 1 (obtained when the SPD is concentrated to the central interval), and the lowest possible score is 0 (when the SPD is concentrated to some other interval).

The quadratic scoring method has the important property of obliging the assessor to make his assessment agree with his true beliefs, i.e., if he wants to maximize his subjective expected score. As long as the partition of the interval [0, 1] is not known to the subject, the method used above for a continuous distribution will have the same property.

3. EXPERIMENTAL PROCEDURE

Subjects

Fourteen students of business administration, with emphasis on quantitative methods, participated in the experiment. Most of them had about one year's study of statistics behind them. The experiment was counted as a seminar within the course requirements, and the students participated on a voluntary basis. No monetary rewards were involved.

The experiment was divided into three sessions of 13 or 14 trials each. The students were divided into two groups. The only difference between the two was that one group received feedback from the earlier sessions. The feedback was given in the form of the true probabilities for the devices in the preceding session. There were also discussions about various related factors such as the degree of uncertainty implied by the quartile ranges and how this uncertainty affected the scores. Eight participants belonged to the feedback group, denoted as Group I. The other group, Group II, consisted of six people.

Devices

Four types of random devices were used. The first type consisted of seven circular discs divided into sectors painted in different colors. The number of sectors and colors varied from 2 to 16 and from 2 to 4, respectively. The discs were meant to be regarded as well-balanced fortune

[3] The choice of the number of intervals and the dividing points are, of course, arbitrary. However, a series of computations of scores were made for a number of different distributions using different partitions of the interval [0, 1]. The dividing points chosen for the experiment were those that seemed to agree best with the experimenter's evaluation of the assessments.

wheels, and the probability to be estimated concerned the event that the wheel would stop at some sector of a specific color.

The second type consisted of eleven noncircular discs divided into irregular fields painted in different colors. Most of these were quadratic, but two were hexagonal and one was a parallelogram. These discs could be regarded as dartboards, and the probability to be estimated concerned the event that a dart thrown from some distance away would hit some field of a specific color (given that the dart hit the board).

The next type consisted of seven proportions like "the proportion of law students at the university." This group of devices differs from the other three in two respects. First, each proportion concerned a characteristic of a finite population, and its true value could be determined exactly. Secondly, all the information was known to the participants before the session, as compared to the other devices where all the information was provided at the session.

Five different kinds of dice made up the fourth type. They were painted in two or three colors. Some were supposed to be rolled and others thrown. Because of their irregular shapes, it was sometimes difficult to tell which surface was "up the most" after a throw. Therefore, the participants were asked for the probability that the die would come to rest on a surface of a specific color.

Determination of the True Probabilities

In order to calculate the score, we have to know the true p values for the various devices. It was easy enough to determine this for the discs by measuring the central angles of the sectors for the circular discs and the areas of the various patches for the noncircular discs. But a long series of throws was the only way to determine p for the dice with sufficient precision. A series of about 1000 throws with each type of die was a reasonable compromise between the demand for precision and the cost of attaining it. One such series was carried out for each type of die.

Presentation

The different types of random devices were presented in mixed order. The first introduced were meant to be easier than the others in order to familiarize the participants with the setup. The degree of difficulty of each type of device was fairly constant thereafter, the only exception perhaps was the noncircular discs, where the ones with the most irregular designs appeared rather late in the experiment. The discs were shown for 3 sec, and the first two dice for 10 sec. The rest of the dice were passed round among the participants for close inspection without a time limit;

In these trials, however, the dice were not to be thrown. The time spent per trial averaged 5 min in the first session but decreased to 3 min in the last two sessions. This time covered the exposure of the device and the time taken to write down the estimates.

Instruction

The instruction period opened the first session and lasted for about 20 min. The subjects did not need a long statistical introduction as they were all familiar with statistical concepts. Most of the time was devoted to familiarizing the subjects with the experimental procedure and the estimates they would be asked to give.

They were asked to determine three points, m, q_1, and q_3, in the following way:

1. Determine a point m such that it is equally probable that p is less than m as that p is greater than m.

2. Assume now that you were told that p in fact is less than m. Conditional on this information determine a new point q_1 such that it is equally probable that p is less than q_1 as that p is greater than q_1 (i.e., is between q_1 and m).

3. (A similar question was asked for the case that p was greater than m. This led to the point q_3.)

The subjects realized that these points were the median and the quartiles, but it was pointed out that this was a useful means of self-interrogation. Some other interpretations of the quantities were also given. After the subjects had fully understood the procedure, it was not felt necessary to repeat it for each trial.

4. ASSESSED DISTRIBUTIONS

It appears that the participants frequently forgot which probability they were asked to estimate and instead estimated some other probability. There clearly must be a mistake when the median 0.13 and the quartiles 0.10 and 0.15 are given for P (yellow) for a yellow, green, and blue disc which is more than 75% yellow. In this and similar cases, it can be argued that "the rules of the game" include keeping track of the instructions, and that corrections or deletions of such answers cannot be permitted. On the other hand, not correcting or deleting them would lead to unusually low scores which would affect further analysis. In seven cases, it seemed obvious that a mistake had been made, and it was decided to delete the corresponding answers.

Very often the three estimates given for a SPD are not compatible with a reasonable probability distribution, let alone one of the β-type. If a probability distribution is to be of the β-type, it is necessary that

$$m - q_1 - (q_3 - m) \begin{matrix} < 0 \\ = 0 \\ > 0 \end{matrix} \quad \text{for} \quad m \begin{matrix} < 0.50 \\ = 0.50. \\ > 0.50 \end{matrix}$$

Table 1 presents the distribution of the answers among the five possible categories. Thirty-nine percent of all answers with the median not equal to 0.50 met the requirement above, whereas, 35% of these answers were symmetric, and 26% were skew in the wrong direction. The number of answers fulfilling the requirement decreased from 40 and 48% in the first two sessions to 31% in the third session. One would rather have expected an increase as the experiment progressed and the participants became better acquainted with the estimation procedure. One explanation of some of the symmetric answers is that almost all estimates were given with a precision of not more than two decimals. It can then happen that the rounded values of the quartiles may well be equidistant from the median.

TABLE 1
CLASSIFICATION OF ANSWERS ACCORDING TO WHETHER THEY COULD BE FITTED TO A β-DISTRIBUTION (FIGURES ROUNDED TO THE NEAREST %)

	Median $\neq$ 0.50			Median = 0.50	
	Right[a]	Symm.	Wrong	Right	Wrong
Group I					
Session 1	35	48	13	4	1
2	54	12	29	6	0
3	31	39	27	3	0
Total	40	33	23	4	0
Group II					
Session 1	43	26	27	1	3
2	32	35	26	6	1
3	28	43	28	0	1
Total	34	35	27	3	2

[a] Right: Answers for which $m - q_1 - (q_3 - m) \begin{smallmatrix} < \\ = \\ > \end{smallmatrix} 0$ when $m \begin{smallmatrix} > \\ = \\ > \end{smallmatrix} 0.50$; Symm.: answers for which $m - q_1 = q_3 - m$ when $m \neq 0.50$; and Wrong: other answers.

In some cases the estimates could be consistent with the subject's beliefs, although no β-distribution could be fitted. This could happen, for example, when the median is rather close to (say less than) 0.50, although it is almost certain that the probability cannot exceed 0.50. The SPD will then be negatively skewed in disagreement with a β-distribution. Still, it is the author's belief that most of these answers were caused by inconsistencies, most of which could probably have been removed if the subjects had been made aware of them. Thus, for example, several sets of estimates imply bimodal distributions, which was probably not what the assessor had intended.

We have described two kinds of answers that do not readily adapt themselves to the β-approximation. The fitted distribution will have roughly the same measures of location and dispersion, but the skewness will be different. Hopefully, the resulting error in the score will not be very large, because the quadratic scoring rule is rather insensitive to small changes in a distribution.

Almost all estimates were given with one or two decimals. This is quite reasonable, of course, since a greater precision would not generally mean anything to the subjects. It is interesting, however, to study the extent to which the subjects used "round numbers" (here, multiples of 0.05) when estimating the median and the quartiles of their SPDs. Table 2 shows the frequencies of the second decimals for the medians and the quartiles. It may be noted that the proportion of "round numbers" for the medians decreased from Session 1 to Session 3 from 65 to 45% for Group I and increased from 49 to 56% for Group II.

TABLE 2
DISTRIBUTION OF SECOND DECIMAL ACCORDING TO THE DIGIT (IN %)

	Medians				Quartiles
	Session				
Digit[a]	1	2	3	1-3	1-3
0	40.6	35.2	24.2	33.1	22.6
5	17.8	17.6	25.8	20.5	20.1
1, 4, 6, 9,	10.0	19.8	14.7	14.8	19.5
2, 3, 7, 8,	24.4	25.3	34.7	28.3	34.9
ODD	7.2	2.2	0.5	3.3	2.9

[a] ODD: Estimates that were either common fractions or were given with three decimals.

5. PRIMARY ANALYSES

The analyses are mainly based on average scores for participant groups, device groups, and sessions. These scores are shown in Table 3. When interpreting them, it should be kept in mind that some of the averages are based on only a few values. Nor are scores for devices with different values of p quite comparable. The same partitioning of the distribution is used for all values of p although there may be some depend-

TABLE 3
AVERAGE SCORES

		Session			
		1	2	3	1–3
Group I					
Device group	1	0.428	0.404	0.466	0.431
	2	0.425	0.447	0.403	0.426
	3	0.398	0.320	0.442	0.395
	4	0.254	0.274	0.390	0.305
	1–4	0.355	0.354	0.416	0.376
Group II					
Device group	1	0.511	0.494	0.439	0.485
	2	0.486	0.436	0.402	0.438
	3	0.366	0.391	0.376	0.377
	4	0.195	0.248	0.269	0.237
	1–4	0.359	0.366	0.353	0.359

ence between the t'ghtness and the location of a distribution. For instance, the distribution will probably be tighter when p is close to zero than when p is around 0.35.

Prior to the experiment it was thought that the "one-dimensional" circular discs would prove easier than the two-dimensional noncircular discs which in turn, would prove easier than the three-dimensional dice. This is also verified by the experimental results. The third device group, the one with the proportions, was the third most difficult.

Group II did not receive any feedback and, therefore, its results reflect the degree of difficulty connected with the various devices in the three sessions. The average score decreased for the circular discs when the number of sectors and colors increased. It also decreased for the noncircular discs when they became more irregular in the second and third sessions. The dice seem to have become easier as the experiment progressed and the participants became acquainted with the different types.

What we are really interested in when we study an individual's performance over the three sessions is whether the third session was his best. Table 4 shows the number of participants who received their highest, medium, or lowest average score in the third session for each device group (and for the total). All eight participants in Group I succeeded best in the third session as compared to only two out of six in Group II. The results of the two groups were rather similar for the first two sessions. In the third session, however, the results of Group I v ere generally superior to those of Group II. The only participant from Group II to split Group I is ranked as number seven. The effect of the feedback seems to be greatest

TABLE 4
Number of Participants Who Obtained Their Highest, Medium, or Lowest Score in the Third Session

	Device group									
	1		2		3		4		1–4	
Result in the third session	I	II[a]	I	II	I	II	I	II	I	II
Highest score	6	1	2	0	4	2	8	4	8	2
Medium score	0	1	2	2	3	1	0	1	0	2
Lowest score	2	4	4	4	1	3	0	1	0	2

[a] I = Group I; II = Group II.

for the dice. This is natural as they were the devices with which the subjects were least familiar and where additional information would have been most valuable.

6. TIGHTNESS OF THE ASSESSED DISTRIBUTIONS

An individual's confidence in his probability estimate of some event is reflected in the tightness of the SPD. This tightness is here measured by the sum of the two parameters a and b, and for ease of reference the sum will be denoted c. A large value of c implies a tight distribution as is evident from the following relation between c and the variance: Var $p = p'(1 - p')/(c + 1)$ where $p' = a/c$ is the mean of the distribution. The median value of c was calculated for each device within each participant group. The average median values for each device group are presented in Table 5.

The results show (i) that the subjects gradually acquired more confidence in their estimates; (ii) that the larger part of this increase can be attributed to Group I, for which the c values were trebled from the first to the third session; and (iii) that the circular discs had the tightest dis-

TABLE 5
Average Median Values of c

	Group I			Group II		
	Session			Session		
	1	2	3	1	2	3
Device group 1	29.9	57.0	93.1	90.8	104.1	79.6
2	20.3	38.4	74.7	53.9	75.4	67.0
3	20.7	16.0	51.3	25.1	23.4	35.8
4	15.8	15.4	51.4	22.3	38.1	61.7

tributions followed by the noncircular discs. The median c value for each device was almost always less for Group I than for Group II in the first two sessions, but this order was completely reversed in the third session. It seems that Group I acquired more confidence after receiving feedback.

The tightness of the SPDs in this experiment can be compared with those in a similar experiment described by Leonardz and Staël von Holstein (1967). Two types of devices were similar in the two experiments, i.e., circular discs and three of the present kinds of dice. In the earlier experiment, the subjects were asked to give estimates of (i) the mean and (ii) a 95% credible interval for their SPDs. A β-distribution was fitted after an approximation of the standard deviation had been obtained by dividing the length of the interval by 4 (the error in this approximation is generally less than 2%, which implies that c was not overestimated by more than 4%). The values of c were three to six times larger than in the present experiment. The average median c value for the discs were 235.2 and for the dice 136.5. The corresponding values for the present study are 68.2 and 31.3 (in the third session they increased to 88.2 and 48.9), respectively.

The comparison shows that the choice of assessment technique can have a substantial effect on the final distribution. The main reason is probably that the subjects find it difficult to fully comprehend the techniques. Similar discrepancies between distributions assessed by different techniques were also found by Winkler (1967a).

7. DISCUSSION

The results of the experiment provide some evidence that it is feasible to train probability assessors. The feedback had some effect, although it was presented only twice. The ideal form of feedback would be immediate feedback after each trial. One can train assessors in several ways, e.g., with feedback in the form of scores, by discussions about the assessments and any inconsistencies, by comparisons of the assessments with actual values, by checking assessments obtained by different methods against each other, etc. The different methods of training will have different effects when the assessment methods are varied, and it would be of interest to study these relationships in greater detail.

The assessment technique in the present experiment was rather simple, mainly because of experimental conditions, and it is highly probable that many of the assessed SPDs did not agree with the assessors' real beliefs. Similar experience led Winkler (1967a) to construct a questionnaire to be used for assessing SPDs. This questionnaire makes use of several techniques. The different assessments can be checked against each other and may ultimately lead to one distribution which, it is hoped, agrees with the assessor's true beliefs. This is naturally a fairly time-consuming procedure when the assessor is rather inexperienced. On the other hand, experiments have now shown that it is necessary to introduce some method by which the assessments can be checked and a questionnaire like Winkler's may prove helpful.

The experiment was restricted to unknown proportions, but future experiments should study other quantities such as parameters from well-known and frequently used distributions. For example, the mean from a normal distribution with known variance or the parameter of an exponential distribution might be chosen. It will, however, be much more difficult to find practical applications that are suitable for such experimentation.

REFERENCES

Edwards, W., Lindman, H., & Savage, L. J. Bayesian statistical inference for psychological research. *Psychological Review*, 1963, **70**, 193–242.

Edwards, W., Phillips, L. D., Hays, W. L., & Goodman, B. C. Probability information processing system: Design and evaluation. *IEEE Transactions on Systems Science and Cybernetics*, 1968, SSC-**4**, 248–265.

de Finetti, B. Does it make sense to speak of 'good probability appraisers'?. In I. J. Good (Gen. ed.), *The scientist speculates: An anthology of partly-baked ideas*. London: Heinemann, 1962, pp. 357–364.

Leonardz, B., & Staël von Holstein, C.-A. S. A comparison between Bayesian and classical methods for estimating unknown probabilities. Project ORBS Technical Report No. 3, Division of Applied Mathematics, Brown University, 1967.

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

Pratt, J. W., Raiffa, H., & Schlaifer, R. *Introduction to statistical decision theory* (preliminary edition). New York: McGraw-Hill, 1965.

Raiffa, H., & Schlaifer, R. *Applied statistical decision theory*. Boston: Division of Research, Graduate School of Business Administration, Harvard University, 1961.

Winkler, R. L. The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical Association*, 1967a, **62**, 776–800.

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

Winkler, R. L., & Murphy, A. H. "Good" probability assessors. *Journal of Applied Meteorology*, 1968, **7**, 751–758.

Received: January 15, 1970