ORGANIZATIONAL BEHAVIOR AND HUMAN PERFORMANCE 3, 36-46 (1968)

Subjective Sampling Distributions and Conservatism[1]

GLORIA WHEELER
University of Michigan

AND

LEE ROY BEACH
University of Washington

When people revise subjective probabilities in light of data, revisions are less than the amount prescribed by the normative model, Bayes's theorem. Previous research suggests that this results from the subjects' lack of understanding of the implications of the data; i.e., from inaccurate subjective sampling distributions. This experiment examined the effects on conservative revisions of training subjects about the implications of data.

The subjects estimated sampling distributions for two binomial populations, were shown samples from the populations in order to teach them veridical distributions, and again estimated sampling distributions. Estimated sampling distributions were good predictors of revisions and, as a result of training, both the sampling distributions and the revisions became more veridical.

A number of recent experiments have shown that when people revise their subjective probabilities about hypotheses in light of data, the revisions tend to be less than the amount prescribed by the appropriate normative model (e.g., Phillips and Edwards, 1966; Phillips, Hays, and Edwards, 1966; Peterson and Miller, 1965; Schum, Goldstein, and Southard, 1966). This phenomenon is called conservatism (Edwards, Lindman, and Phillips, 1965) and two explanations have been advanced to account for it. The first is that subjects have difficulty in performing the mechanics of integrating the meaning of the data into their subjective

[1] This research was performed at the Engineering Psychology Laboratory, Institute of Science and Technology, The University of Michigan. The research was supported by USPHS Fellowship MF-12,744, USPHS grant NIH-MF-11496, both from the National Institutes of Mental Health, and by National Institute of General Medical Sciences grant 13143-01. The authors thank Drs. Cameron Peterson and Ward Edwards for their help and advice.

probabilities (Phillips, 1966). However, the orderliness of conservative revisions, together with evidence that subjects can perform complex and subtle subjective manipulations of their subjective probabilities (Beach, 1966; Beach and Peterson, 1966; Peterson, Ulehla, Miller, Bourne, and Stilson, 1965), suggests that the cause may lie somewhat deeper than arithmetic errors. The second explanation is that subjects may not always understand the implications of the data for the hypotheses under consideration. This is especially plausible in the extremely abstract experimental tasks in which conservatism has been found.

The usual paradigm is to tell the subjects that there are two urns (the hypotheses) full of poker chips. One urn contains, say, .70 red chips and .30 blue chips and the other contains .30 red chips and .70 blue. Out of the subjects' sight, the experimenter flips a coin to select an urn, draws a random sample of chips with replacement, and displays the sample. The subjects' task is to estimate the probability of each urn having been the source of the observed sample. To perform this task, the subjects' initial subjective probabilities of .50–.50 must be revised in light of the data and the revised subjective probabilities stated as estimates. This is the procedure a statistician would follow using Bayes's theorem, the appropriate normative model (Edwards *et al.*, 1965), and the research to date shows that this is generally the same procedure that subjects follow. However, suppose that the sample contained eight red chips and four blue ones. Using Bayes's theorem, the statistician would state a revised probability of .97 for the .70 red urn and of .03 for the .70 blue urn, while subjects typically estimate about .75 and .25, respectively; subjects do not revise as much as the statistician does for the same data. Or, stated differently, subjects apparently fail to regard the data as being as diagnostic as the statistician does.

For a statistician the diagnosticity of a sample is given by the sampling distributions for the hypotheses under consideration, in this case for .70–.30 binomial populations. So, to say that subjects regard data as less diagnostic than the statistician does is to say, formally, that their sampling distributions are more like rectangular distributions than theoretical distributions are.

When looked at in this way it is clear that investigation of the second explanation of conservatism, that subjects do not understand the implications of the data, involves the investigation of subjective sampling distributions. The subject's estimates of sampling distributions should be too flat. Moreover, the errors in the estimated sampling distribtuions should correspond to the errors in the revisions that presumably rely on the inaccurate distributions. And finally, if, through training, subjective sampling distributions could be made more accurate, and if revisions became cor-

respondingly more optimal, it would provide supporting evidence for the hypothesis.

This investigation was performed by Peterson, DuCharme, and Edwards (1967) using the urn and poker chip paradigm described above. Subjects revised their subjective probabilities for a number of samples that were each preceded by the random selection of an urn and a random draw with replacement. After this they estimated the sampling distributions for the urns. Then, to teach the subjects what the theoretical distributions are like, successive samples were drawn from an urn and displayed by being tallied in a frequency distribution. As the samples were drawn the subjects revised their own estimates of the sampling distribution until, at the end of training, their estimates looked like the theoretical distribution. Then they made a second set of subjective probability revisions for the samples drawn from the randomly selected urns.

It was found that the initial revised probabilities were conservative and that estimated distributions were initially too flat. By using the values from these estimated sampling distributions in Bayes's theorem instead of values from the theoretical distributions, it was possible to predict the subjects' conservative revised probabilities. This appears to be strong evidence that subjective revision procedures and Bayesian procedures are the same and that the subjects' flat sampling distributions are responsible for conservatism. However, the second, post-training, set of revised subjective probabilities was only slightly less conservative than the first. The process of teaching the subjects about sampling distributions apparently had very little effect.

The conservatism of the second set of revisions mitigates the evidence provided by the consistency between the first set and subjects' flat estimated sampling distributions. It is possible that in this experimental setting subjects tend to hedge their bet by flattening all of their response distributions, be they revisions, estimated sampling distributions, or whatever. If so, the flat distributions could be consistent with the conservative revisions even though the sampling distribution hypothesis is wrong. Until a strong link between conservatism and subjective sampling distributions can be shown, preferably by commensurate changes in both as a function of training, the consistency results must remain, at best, circumstantial and inconclusive evidence for the hypothesis. The purpose of this experiment was to investigate the effects of training and by doing so to demonstrate the link between revision of subjective probabilities and subjective sampling distributions.

There are two salient possibilities that might have caused the second set of revisions to be conservative. The first is that the training procedure had no effect and the second is that it was effective but that the subjects

were attempting to decrease the risk of gross error by making conservative responses in the second revision task. In the experiment to be reported, the training method was changed and payoffs were added to the revision task to discourage distortion of responses.

The training method was similar to the one used by Peterson *et al.* (1967) except that it perhaps was more compatible with subjects' normal methods of learning about the implications of data for hypotheses. The usual course of subjective inference is from observations of data to statements about hypotheses, i.e., the focus is on the probabilities of various hypotheses given the observed data. Sampling distributions involve just the opposite kind of probability statements, the probability of obtaining the observed data assuming a particular hypothesis to be true. While subjective sampling distributions are basic to revisions and inferences, they are seldom the central focus of subjects' thinking about the task; they are formed by incidental learning in the course of observing data in relation to hypotheses. Therefore, instead of concentrating their attention on sampling distributions in order to increase verdicality, the subjects saw samples, made bets about which population was the source of the samples, and then were told which population was correct. The distribution of displayed samples conformed to the theoretical sampling distributions for the two urns, betting focused on the probabilities of each of the two urns given each sample, and feedback provided the necessary information for learning about the relation between the samples and the urns, i.e., about sampling distributions. The subjects estimated the sampling distributions at the beginning and end of training and these estimates were compared with the subjective probability revisions inferred from the bets made early and late in training to see if they were consistent both times and if inaccuracy and conservatism had decreased with training.

METHOD

Apparatus

The two urns were represented by two cards with ten poker chips on them. One urn contained .80 red chips and .20 blue; the other urn contained .40 red chips and .60 blue. In addition, nine other cards, each with eight poker chips on them, were used to display the nine possible samples for a sample size of eight. One card had eight red chips and zero blue chips, one had seven red chips and one blue, etc.

Subjective binomial distributions were obtained by having the subjects distribute 100 markers among a set of nine troughs; each marker represented .01 and each trough represented one of the possible samples.

The markers were distributed roughly equally to begin with and the subjects modified the arrangement to represent their estimates. Estimates of a fraction of .01 were permitted, but very few were made.

Bets were made by selecting the pair of payoffs for which the subject was willing to bet in light of his subjective probabilities for the urns given the particular sample that had been drawn. A list of 51 pairs of payoffs was printed on the pages of an answer book and, for each sample, the subjects indicated which urn they thought was most likely and then marked the pair of payoffs for which they were willing to bet.

Payoffs were in the form of points and the list conformed to Toda's (1963) quadratic payoff function. If a subject attempted to maximize his subjectively expected winnings in points, only one bet on the list was optimal in view of his subjective probabilities given the observed sample. By assuming that subjects do attempt to maximize their earnings it is possible to reason backwards from their choice of payoffs and to infer the subjective probabilities that underlie the choice. This assumption is sound because the subjects were told at the start that their pay for participating in the experiment was entirely dependent upon the number of points accumulated in the course of training. (For a more detailed explanation of the betting and payoff procedure see Beach and Phillips, 1967.)

Procedure

First, the subjects were told about the urns and the sampling procedure. Then they estimated the probabilities of obtaining each of the nine kinds of samples on a random draw from the $P = .80$ red urn and from the $P = .60$ blue urn, i.e., they estimated the sampling distributions.

Then the betting procedure was explained and the subjects were told that to help them better understand the relation of the samples and the urns, they would be shown a series of random samples of chips and that the sampled urn had been selected by flipping a fair coin before each sample was drawn. After seeing a sample they were to bet on which urn it was drawn from by indicating the urn they favored and selecting a pair of payoffs from the list. If the urn that they bet on actually was the source of the sample they would receive the high payoff in the pair, otherwise they would receive the low payoff. After they made their bets the experimenter told them which urn was correct and each subject added the appropriate number of points to his running total on an adding machine that sat on the desk in front of him. The procedural sequence was this: The subjects estimated sampling distributions, then saw and made bets on 100 samples, the first 20 of which were used for the revision-task data analyses. Then they reestimated the sampling distributions

(primarily to relieve boredom), saw 100 more samples, and made their final set of sampling distribution estimates. Then they saw a final 20 samples that were identical to the first 20 samples and that also were used for revision-task analyses.

Subjects. Seventeen male university students served as volunteer subjects in groups of 3, 4, and 5 at a time.

RESULTS

In Fig. 1 the subjects' estimated sampling distributions are compared with the theoretical distributions. The estimated distributions were obtained by computing each subject's likelihood ratio for each sample; the quotient of his estimate of the probability of the predominantly red urn given the sample and his estimate of the probability of the predominantly blue urn given the sample. Then median likelihood ratios were obtained across subjects for each sample. The probabilities plotted in Fig. 1 are the two estimated probabilities that comprised each of these median ratios (normalization was required to make the distributions total 1.00 but the change in the points was slight).

The subjects' first estimates of the sampling distributions were too flat compared with the theoretical distributions. In general, the training procedure brought them into closer correspondence with the theoretical values, although complete accuracy was never attained. Indeed, for some samples the subjects' distributions became more extreme than the theoretical distributions. The distributions estimated in the middle of training fall between the curves in Fig. 1 but, for the sake of clarity, they are not presented.

The degree of correspondence between the estimated distributions and the theoretical distributions can be illustrated better by plotting the logs

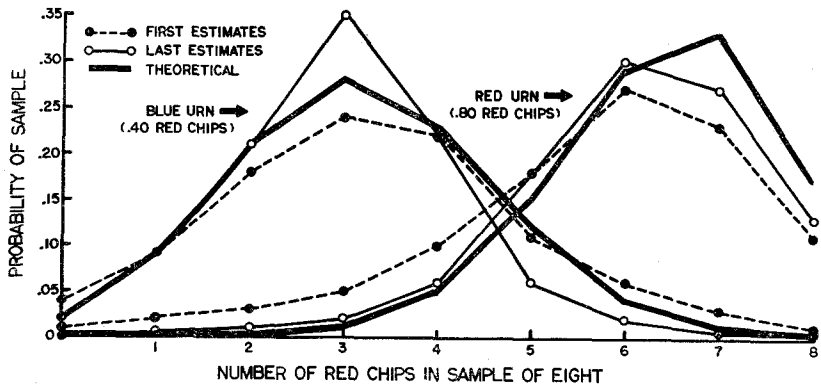

FIG. 1. Estimated sampling distributions before and after training.

of the median likelihood ratios underlying the subjective curves in Fig. 1 against the logs of likelihood ratios obtained from the theoretical curves.[2] This comparison, Fig. 2a, clearly shows the conservative nature of the first estimates. Training brought the subjective likelihood ratios, i.e., subjective sampling distributions, into line with the theoretical values although there was a bias toward favoring the predominantly red urn. This bias was found throughout the data and will be discussed later.

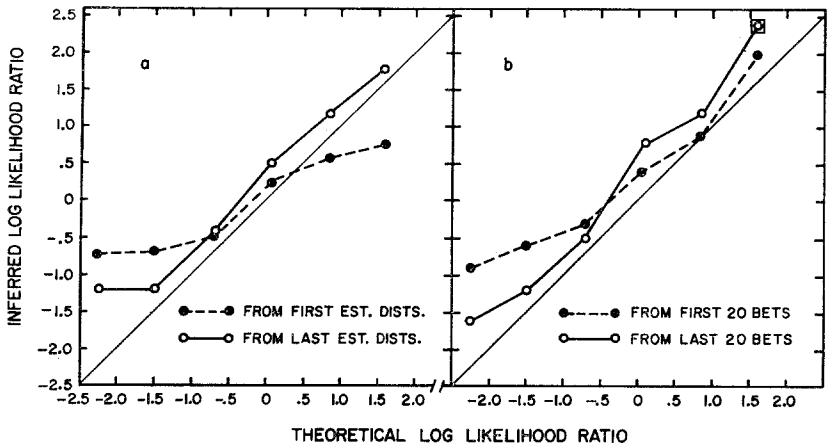

FIG. 2. Log likelihood ratios inferred (a) from estimated sampling distributions and (b) from revision responses (bets) compared with theoretical values before and after training.

The curves in Fig. 2a are quite representative of the curves for individual subjects. This is illustrated best by the correlations and regression slopes, for each subject, between the theoretical log likelihood ratios and the log likelihood ratios constructed from his estimated distributions. For the median values in Fig. 2a, the correlations were .95 before training and .93 afterwards. More important, however, are the slopes of the regression lines for predicting the subjective ratios from the theoretical ratios. These were .44 before training, indicating flat subjective sampling distributions, and .85 after training, indicating more verdical distributions. The analysis for individual subjects yielded median correlations of .92 (with 53%, i.e., 9, of the 17 subjects' correlations lying between .90 and .95) before training and .91 (53% between .87 and .95) after training. The median slope was .45 (53% between .37 and .55) before training,

[2] Minor irregularities in the very small probabilities for samples of 0 red-8 blue, 1 red-7 blue, and 8 red-0 blue lead to extremely deviant likelihood ratios. In addition, these samples seldom occurred during training and yielded very little and very unstable data. Because both factors lead to unrepresentative results, data for these points were deleted from all analyses.

indicating flat distributions and .96 (53% between .60 and 1.11) after training, indicating increased veridicality.

Thus far both the group results in Fig. 2a and the individual subjects' results conform to those reported by Peterson *et al.* (1967). The next question is whether the training used in this experiment successfully influenced subjective probability revisions. A comparison similar to that in Fig. 2a is afforded by data from the 20 identical bets at the beginning and end of training. The ratio form of Bayes's theorem is

$$\frac{P(U_r|s)}{P(U_b|s)} = \frac{P(s|U_r)}{P(s|U_b)} \cdot \frac{P(U_r)}{P(U_b)},$$

where U_r and U_b are the predominantly red and predominantly blue urns respectively and s is the observed sample of chips. In this experiment the probabilities prior to the presentation of each sample were always $P(U_r) = P(U_b) = .50$ and their ratio was 1.00. The posterior probabilities, $P(U_r|s)$ and $P(U_b|s)$, were inferred from the subjects' bets and, because the ratio of the prior probabilities is 1.00, the ratios of the posterior probabilities could be assumed equal to the likelihood ratios,

$$P(s|U_r)/P(s|U_b).$$

This permitted the subjects' likelihood ratios for each of the 20 trials at the beginning and end of training to be constructed from the subjective probabilities inferred from their choices of bets.

Figure 2b shows the relation between the median log likelihood ratios inferred from both the first and the last bets and the corresponding theoretical log likelihood ratios.[3] The correlations between these variables for the data in the figure were .86 before training and .93 after training. The respective slopes were .72, indicating conservatism at the beginning of training, and 1.05, indicating accurate revisions at the end of training. Analyses for the individual subjects yielded median correlations at the beginning and end of training of .86 and .89 (53% between .76 and .90 and between .86 and .92, respectively). The median slopes were .70 (53% between .45 and .97), indicating conservatism at the beginning, and 1.13 (53% between .84 and 1.45), indicating more nearly accurate revisions at the end of training.

Comparison of Figs. 2a and 2b and the results for individual subjects, shows that both the distribution estimates and the bets (revisions) yielded likelihood ratios that were less conservative after training. The

[3] The two boxed points on Figs. 2b and 3 indicate median inferred probabilities of 1.00 from the bets. There is no ratio for 1.00 and .00 and so it was assumed that the subjects were trying to indicate a probability slightly higher than the next lowest probability, and consequently that value, .995, was used.

final question is whether the two sets of ratios, those inferred from the estimates and those inferred from the bets, both reflect a common set of subjective sampling distributions. The comparison is shown in Fig. 3. Except for the deviations at the extreme ratios, which will be discussed directly, the two kinds of ratios are quite similar. For the group data in Fig. 3, the correlations between the log likelihood ratios from the esti-

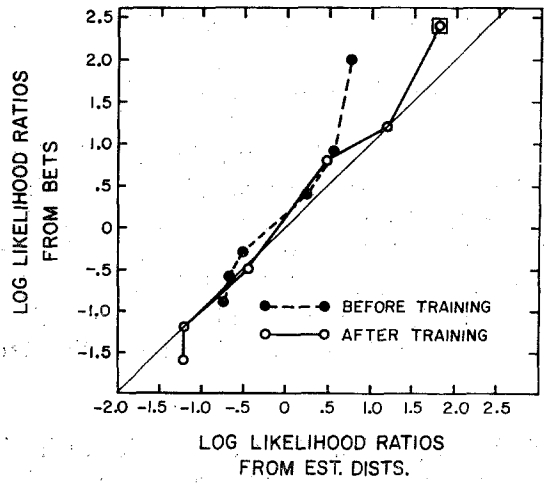

FIG. 3. Consistency of inferred log likelihood ratios from estimated distributions and from revision responses (bets) before and after training.

mated distributions and from bets before and after training are .90 and .99 with slopes of 1.57 and 1.22. The analysis of the individual subjects' data gave a median correlation of .85 (53% between .72 and .93) before training and .94 (53% between .90 and .97) after training. The median slope was 1.27 (53% between .76 and 1.93) before training and 1.29 (53% between .90 and 1.57) after training.

The high slopes show that the log likelihood ratios inferred from bets (revisions) were slightly more extreme than those inferred from the estimated distributions. This is not surprising because, as Beach and Phillips (1967) found, this betting method yields values more extreme than estimated values for very high and very low estimated probabilities. It is these extreme probabilities that constitute the large and small likelihood ratios and produce the distortions illustrated in Fig. 3 and found in the individual subjects' data.

The systematic distortion of the slopes in this comparison must not be allowed to cloud the more fundamental finding that the relation between the two sets of ratios changes very little from the beginning to the end of training. What change there was consisted primarily of decreased re-

sponse variance, as evidenced by increased correlations, and a corresponding decrease in the variance of the distribution of slope values. These changes are probably attributable to the subjects' increased familarity with the tasks as training progressed. Beach and Phillips (1967) showed that the stability of subjects' probability estimates and bets, and therefore correlations involving them, increased in the course of training.

One aspect of the results, mentioned above, demands further comment. Thoughout both the group results in the figures and the individual subjects' results, there was a fairly consistent bias in favor of the predominantly red urn. This may have resulted from the asymmetry of the two urn compositions and the consequent asymmetry of the sampling distributions. In the usual subjective probability-revision experiment the two urns are symmetric, so that, for instance, a sample of 2 red-6 blue chips would favor the predominantly blue urn by exactly the same amount that a 6 red-2 blue sample would favor the predominantly red urn. In this asymmetric case, of the nine possible samples, five favor the blue urn and four favor the red urn, with one of the four just slightly favoring the red urn (Fig. 1). The subjects appear to have been influenced by this, imposed symmetry on the samples, and thereby were biased toward the red urn. The interesting thing however, is that this bias appeared in both the estimates and in the bets (revisions) further support for the hypothesis that both reflect the subjects' underlying subjective sampling distributions.

In summary, the results demonstrate the link between conservatism and the inaccuracy of subjective sampling distributions. At the beginning of the experiment subjects' estimated sampling distributions were too flat, their revisions were conservative, and the two errors were consistent with one another. At the end of training the estimated distributions were more veridical, the revisions were less conservative, and the two sets of responses were still consistent with one another. The effects of training and the consistency between estimates and revisions at the end of training go beyond the Peterson *et al.* (1967) results and provide additional evidence that conservatism is due at least in part to subjects' failure to understand the implications of the data for the hypotheses under consideration.

REFERENCES

BEACH, L. R. Accuracy and consistency in the revision of subjective probabilities, *IEEE Transactions on Human Factors in Electronics,* 1966, *HFE-7*, 29–37.

BEACH, L. R., AND PETERSON, C. R. Subjective probabilities for unions of events. *Psychonomic Science,* 1966, **5**, 307–308.

BEACH, L. R., AND PHILLIPS, L. D. Subjective probabilities inferred from estimates and bets. *Journal of Experimental Psychology,* 1967 in press.

EDWARDS, W., LINDMAN, H., AND PHILLIPS, L. D. Emerging technologies for making decisions. *In* Newcomb, T. (Ed.), *New directions in psychology.* II. New York: Holt, Rinehart, and Winston, 1965. Pp. 259–325.

PETERSON, C. R., DUCHARME, W. M., AND EDWARDS, W. Sampling distributions and probability revisions. *Journal of Experimental Psychology* (in press).

PETERSON, C. R., AND MILLER, A. J. Sensitivity of subjective probability revision. *Journal of Experimental Psychology,* 1965, **70,** 117–121.

PETERSON, C. R., ULEHLA, Z. J., MILLER, A. J., BOURNE, L. E., AND STILSON, D. W. Internal consistency of subjective probabilities. *Journal of Experimental Psychology,* 1965, **70,** 526–533.

PHILLIPS, L. D. Some components of probabilistic inference. Human Performance Center, University of Michigan, Technical Report No. 1, 1966.

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

PHILLIPS, L. D., HAYS, W. L., AND EDWARDS, W. Conservatism in complex probabilistic inference. *IEEE Transactions on Human Factors in Electronics,* 1966, *HFE-7,* 7–18.

SCHUM, D. A., GOLDSTEIN, I. L., AND SOUTHARD, J. F. Research on a simulated Bayesian information-processing system. *IEEE Transactions on Human Factors in Electronics,* 1966, *HFE-7,* 37–48.

TODA, M. Measurement of subjective probability distribution. Division of Mathematical Psychology, Institute for Research, State College, Pennsylvania, University Park, Report 3, April, 1963.

RECEIVED: April 10, 1967