ORGANIZATIONAL BEHAVIOR AND HUMAN PERFORMANCE 10, 364–373 (1973)

The Best Guess Hypothesis in Multistage Inference[1]

CHARLES F. GETTYS[2]
University of Oklahoma

CLINTON KELLY III AND CAMERON R. PETERSON[3]
University of Michigan

Intuitive multistage inferences are typically excessive when compared with the optimal model, a modified form of Bayes' theorem. One explanation for this excessiveness is that the *S* primarily attends to the implications of the probable event described by the first-stage inference, neglecting the implications of less likely events. If a *S* follows this strategy, called a "best guess" strategy, then a testable implication is that his probability revision at the upper level should be insensitive to variations in the distribution of probabilities across all but the most likely event described by the first-stage inference. The results of the present experiment support this hypothesis.

Multistage inference consists of a series of single-stage inferences where the output of each previous stage becomes the input to the next stage. In a single-stage inference men reason from data or unambiguously observed evidence to a set of hypotheses. Multistage inference starts with the same unambiguous data or evidence in the first stage; however, the input for the next stage is the output of the previous stage. The next stage of inference is therefore based on the probabilities of events, rather than upon definite knowledge that a particular event is true (Gettys & Willke, 1969).

For example, suppose you wanted to predict the success or failure of a large garden party. Assume that the party is less likely to be success-

[1] This study was conducted while the senior author was an NSF Post-doctoral Fellow in the Engineering Psychology Laboratory at the University of Michigan. This paper was supported in part by the National Aeronautics and Space Administration Grant No. NGL 23-005-171 and the Advanced Research Projects Agency and the Office of Naval Research Grant Nos. NONR-N-00014-73C-0149 and NR-197-023.

[2] Requests for reprints should be sent to Charles F. Gettys, Department of Psychology, University of Oklahoma, Norman, Oklahoma 73069.

[3] Now at Decisions and Designs, Incorporated, Suite 600, 7900 Westpark Drive, McLean, Virginia 22101.

ful if it is crowded indoors because of rain. Your datum is the presence of a dark cloud on the horizon. The first stage of inference would relate the dark cloud to the presence or absence of rain during the party. Suppose you estimated that the probability of rain was .70. This estimate would become the input to the next stage of inference. If you knew with certainty that it would rain, then you could infer the probability that the party would be a success. But you are not entirely sure that it will rain; the data that you have indicates rain with a probability of .70, so how should you proceed?

Modified Bayes Theorem (MBT) provides an optimal model for such multistage inferences (Gettys & Willke, 1969; Dodson, 1961). A number of studies have shown that intuitive performance in a multi-stage task results in *more* certainty being extracted from the data than is predicted by the MBT model. For example, in an odds estimation task the *S*s' odds are typically larger than those calculated by MBT. This result is quite surprising because evidence indicates human performance in a single-stage inference task is almost always conservative; i.e., humans extract less certainty than warranted by the data (e.g., Edwards, 1966). The paradox, of course, is that a multistage inference is a series of single-stage inferences. If people extract less certainty than the data warrant in single-stage inferences, then in the multistage situation one might expect the *S*s to become more and more conservative with each succeeding stage since their departures from nonoptimality should accumulate from stage to stage. In fact the reverse is true; *S*s are more certain at the end of two stages of inference than is warranted by the optimal model (MBT). This suggests that some process occurs at the "interface" of the single stage tasks which is so excessive that any single-stage conservatism is overcome.

The single-stage inference task is always based upon data which are known with certainty. However, even though a multistage task starts with certain data, succeeding stages of inference deal with uncertain data. Several models have been formulated to explain how having to deal with the probabilities of data instead of certain data might create excessive certainty in multistage inference. One nonoptimal model having the property of predicting excessive certainty is the "As-If" model (Gettys and Willke, 1969; Howell, Gettys, and Martin, 1971). This model, designed for situations where people have the option to collect more data if they feel it is needed, assumes that data collection continues in the first stage of inference until the decision maker is sufficiently sure of the state of the world. Once his certainty exceeds some threshold value, he then proceeds to the next stage of inference, acting "as-if" he were entirely certain of the input to the next stage. To return to the

garden party example, the decision maker, after seeing the dark cloud, might get a current weather report. Suppose a severe storm warning were forecast. His certainty for rain probably now would exceed his threshold value, and he would proceed to the second stage of inference acting "as-if" he were certain of rain. The result of the second stage of inference would be his estimate of the probabilities of success or failure based on his "as-if" assumption of rain. *His assessed probability for failure should now exceed the veridical (MBT) probability for failure because by making the "as-if" assumption of rain he is ignoring the possibility that it might not rain.* If, in fact, his "as-if" assumption is incorrect and it doesn't rain, then the party probably will be a success. The optimal model considers both possibilities, rain and no rain, in assigning probabilities to success or failure. The "as-if" model considers only the possibility of rain, and for this reason leads to excessive certainty that the party will be a failure.

How might a person behave if his certainty about the input to the second stage of inference were less than the threshold value required for an "As-If" assumption and there were no hope of increasing his certainty with more data? One possible hypothesis that is consistent with the excessive certainty found in previous studies is that he will first make an "As-If" assumption that is at best a guess. This model, termed the "Best Guess" model, is in effect a qualified "As-If" model and shares with the "As-If" model the idea that the decision maker will either ignore or tend to ignore the implications of the other less-likely events in the second stage of inference by concentrating almost exclusively on the most likely event. In terms of the example, if the only information you have is the dark cloud on the horizon, you might not be willing to make an *unqualified* "As-If" assumption, but you might first assume that it is going to rain and arrive at subjective odds for success based on this assumption. Then because you are not entirely certain that it will rain, you might reduce your subjective odds somewhat to take this into account. These subjective odds might well be different from those calculated with MBT, primarily because you have not explicitly considered the implications of no rain.

Snapper and Fryback (1971) reported results which are consistent with the above explanation in an experiment concerned with data reliability. However, their procedure did not permit a direct test of the Best Guess Model; that is the purpose of the present experiment.

METHOD

The goal of the experiment required at least three levels of variables constructed in such a manner that the intermediate level variable con-

tained more than two events. It further required a manipulation of the probability distribution across all but the most likely of the intermediate events—a manipulation that would have a resulting impact on the magnitude of optimal probability revision at the upper level as the result of the occurrence of an event at the lower level.

Consequently, the three levels took the following form. The upper-level variable was comprised of two bags labeled I and II, respectively. Each bag served as a container that was filled with smaller containers which represented intermediate-level events. Specifically, each bag contained 18 small cans (35 mm. film cans) and each can was labeled with either *A, B, C,* or *D*. Finally, each can contained 100 small colored discs; each disc was either red, green, yellow or blue.

The composition of each container is described in Table 1. Part A of the table describes the bag composition with respect to cans and Part B of the table describes the can composition with respect to disks. For example, 8 cans labeled *A* are in Bag I whereas only 1 can labeled *A* is in Bag II. As shown in Part B, 80 discs are in Can *A*, 1 in Can *B*, 1 in Can *C*, and 18 in Can *D*.

The experiment proceeded as follows. One of the two bags was selected at random, a can was sampled at random from that bag, and a disc was sampled at random from that can. Thus, the draw of a red disc provides

TABLE 1

NUMERICAL COMPOSITION OF BAG AND FILM CAN COMPONENTS

A

BOOKBAG COMPOSITION

CAN LETTERS	BOOKBAG I	BOOKBAG II
A	8	1
B	3	6
C	6	3
D	1	8

B

FILM CAN COMPOSITION

DOT COLORS	CAN A	CAN B	CAN C	CAN D
RED	80	1	1	18
GREEN	1	80	18	1
YELLOW	18	1	80	1
BLUE	1	18	1	80

evidence in favor of Can *A*, which in turn provides evidence in favor of Bag I. Notice that it is only the bottom level event, a disc, that is directly observed. That observation provides only partial evidence with regard to the intermediate-level event, the can, which in turn provides partial evidence about which upper level event was selected. Thus, the first stage of inference relates disc color to can letter and the second stage of inference relates can letter to bag number.

The strategy of acting as if the most likely event is true at one level will lead to probability distributions that are extreme at the next higher level. Thus, this strategy is consistent with the empirical result that people revise upper-level probabilities excessively at a multistage task.

There is another testable hypothesis that can be derived from the best-guess strategy. If a person acts as if the most likely event is true at any intermediate level, he then ignores the probability distribution across all other events at this level. His probability revision at the upper level should therefore be insensitive to variations in the distribution of probabilities across all but the most likely event at the intermediate level. The present experiment was designed to test that hypothesis.

Experimental Design

Three inference tasks of the type shown in Table 1 were constructed. The frequencies shown in Part A were used in all three tasks. The matrix shown in Part B was used in one task; in the other two tasks the value of 80 in the lower matrix was changed to either 70 or 90, and the value of 18 was changed to either 28 or 8, respectively. For purposes of later discussion these three tasks will be designated as the 70-28, the 80-18, or the 90-8 task. In all three tasks the *Ss* estimated the odds of the bags given the color of a single disc drawn from the can.

Subjects

The 25 *Ss* were University of Michigan students who had previously served in another multistage inference experiment lasting about two hours. In the previous experiment *Ss* had been trained in the response mode required, and had made an extensive series of odds estimates in a multistage inference task. However, the optimal model was never discussed, nor was any type of feedback used.

Instructions to Ss

The instructions were brief because of the previous experience of the *Ss*. The details of the task were explained. The *Ss* were asked to imagine that a bag had been randomly selected on the basis of a toss of a fair

coin, that a can was then randomly drawn from the bag, and that a paper disc was randomly drawn from the can. Then they were asked to assume that a disc of a particular color was, in fact, drawn according to this random process, and were asked to estimate the odds of the bags on the basis of the color of the disc.

Procedure

Following the instructions, the *Ss* estimated the odds of the bags in all three tasks. Matrices like those in Fig. 1 were used to inform the *Ss* of the relative frequencies of the cans and the discs. The tasks were presented in a random order for each group of 4 to 6 *Ss*. Within each task each of the 4 possible colors was used in random order. The *Ss* estimated the odds of the bags for all possible colors before moving to the next task. When the *Ss* had completed the twelve estimates (4 colors per task × 3 tasks), the three tasks were then repeated using different random orders for a total of 24 judgments, two for each color in each task.

RESULTS AND DISCUSSION

An inspection of the data showed an extreme bimodality in the *Ss*' odds responses. For some *Ss* the theoretical difference between the blue and red dots, and the difference between the yellow and green dots, caused no difference in the odds estimates. Other *Ss* were more extreme in their odds estimates with a blue dot than they were with a red dot, and more extreme with a yellow dot than with a green dot. These latter *Ss* were consistent with MBT in at least an ordinal sense. It appeared that some subjects were "unaware" of the blue–red and the yellow–green differences, while other *Ss* were "aware" to the extent that they were responding in at least the right direction. With this thought in mind, all *Ss* who responded with at least one odds estimate for blue that was at least 2% greater than the odds estimate for red, or an estimate for yellow that was at least 2% greater than green, were classified as "aware" *Ss*. These *Ss* were at least marginally "aware" because for one or more judgments their odds estimates changed in the blue–red pair and the yellow–green pair in the direction that MBT dictates. Ten *Ss* of the 25 were classified as "aware" *Ss* by this conservative criterion.

The other 15 *Ss*, the "unaware" *Ss*, showed no tendency to respond differently to changes in the probabilities of the less likely events. They literally ignored the implications of the less likely cans. Their responses are consistent with an extreme form of the "Best-Guess" model. The medians of the responses of the "unaware" *Ss* are shown in Fig. 1. Because the bag that the odds favor is formally irrelevant, the data are plotted on an absolute log odds scale. The median log odds responses

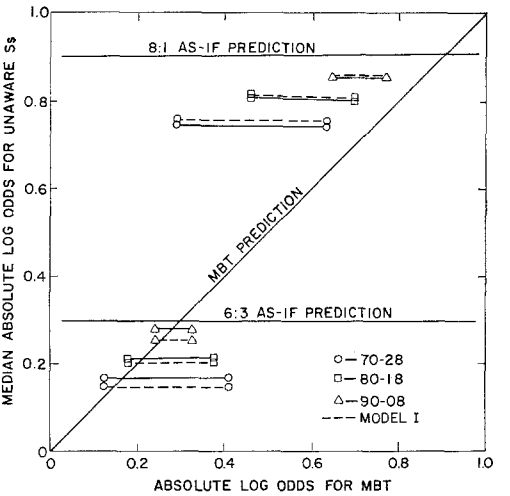

FIG. 1. Best-guess, as-if, and MBT models as predictors of performance of "unaware" *Ss*.

to the red and the blue discs are connected by a solid line in the upper part of the figure for the three levels of data uncertainty, and the medians for yellow and green discs are similarly jointed in the lower part of the figure. Also shown in Fig. 1 are predictions for MBT (the line on the positive diagonal), predictions for the "As-If" model (the two horizontal lines) and predictions for a version of the "Best Guess" model, termed Model I in the figures.

The model I predictions are obtained by multiplying the probability of the most likely event by the posterior odds obtained if that event were true. Suppose a red disc were drawn in the 80-18 task. The probability of Can *A* is .80 and the odds are 8:1 if in fact the dot came from *A*. The Model I prediction would then be $0.8 \times 8/1 = 6.4$ or odds of 6.4:1. The "As-If" model predicts odds of 8:1 for the blue and red discs and 6:3 for the yellow and green dots *provided that the threshold certainty for can type is exceeded.* For MBT the optimal odds for a red dot are 2.86:1, and may be calculated according to the following formula for the posterior odds (adapted from formula 5 in Gettys and Willke, 1969):

$$\frac{P(BI|color)}{P(BII|color)} = \frac{P(BI)}{P(BII)} \times \frac{\sum_i P(color|can_i)\ P(Can_i|BI_I)}{\sum_i P(color|can_i)\ P(Can_i|BI)}, \qquad [1]$$

where B stands for bag, and the other entries are calculated from conditional probabilities such as shown in Table 1.

The data in Fig. 1 are clearly not fitted by either the "As-If" or the MBT predictions. The *Ss* responses are less extreme than the "As-If" predictions for the upper blue–red pairs, where the "As-If" prediction is 8:1 odds, and are similarly less extreme than the 6:3 odds prediction in the lower part of the figure. However, the extreme version of the "Best Guess" Model, Model I, fits the Fig. 1 medians very well. The horizontal dashed lines in Fig. 1 are the Model I predictions. For all tasks, the Model I predictions are to the right of the MBT diagonal for the yellow dots in the lower part of Fig. 1. The "As-If" model and Model I do not necessarily predict odds estimates that are more extreme than MBT odds. These points arise, for example, in the 80-8 task when Can *C* is most likely (P = .80) and can A (P = .18) is less likely. The most likely event gives 6:3 odds for Bag I if true and the less likely event gives odds of 8:1 for Bag I. In this case, any model which ignores the 8:1 ratio furnished by the less likely event will be conservative in respect to MBT.

If it is assumed that *Ss* will not adopt a nonoptimal model if it deviates too much from their subjective feeling of certainty, then perhaps the important result is that *Ss* used Model I because they saw nothing wrong with it. The magnitude of their odds response was determined by Model I but in another situation they might use some other combination rule. More importantly, the fact that Model I predictions do fit the data suggests that *Ss* tended to concentrate on the most likely alternative, and ignored the implications of the less likely alternatives.

The data for the 10 "aware" *Ss* are presented in Fig. 2. As in Fig. 1, the predictions of the "As-If" and the MBT models are shown in the figure, but the Model I predictions are omitted because they clearly do not fit the data.

In general, the "aware" *Ss* seem to respond to the same variables as MBT, but the quantitative fit of the MBT model is poor. *Ss* are characteristically more certain than the MBT model, as has been found in previous research. Like MBT, the *Ss* are less certain than implied by the "As-If" model for the blue and the red discs. Also, as in MBT, their judgments to the yellow disc exceeds the "As-If" prediction. This, of course, occurs because the most likely event has odds of 6:3 and the less likely event has odds of 8:1. If the *Ss* are aware of the nuances of the multistage situation, they should realize that the odds must be greater than 6:3. The only exception to this general picture is the location of the 80-18 data for the yellow and green discs. The posterior based on yellow odds should increase as the probability of the most likely

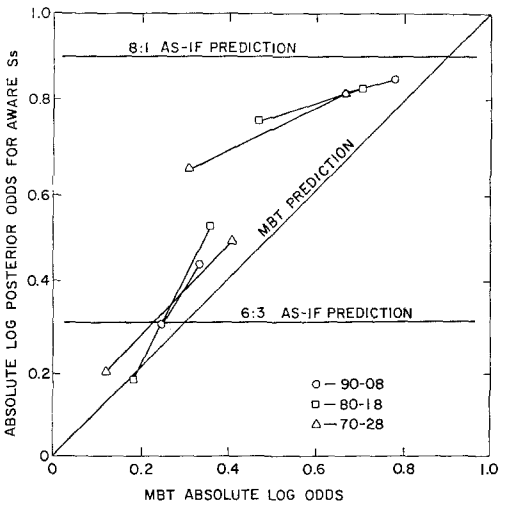

FIG. 2. Performance of aware *Ss* compared to MBT and as-if.

event decreases, while the odds based on green should decrease as the probability increases. The responses in the 80:18 condition do not follow this pattern. In general, the "aware" *Ss* seem to be using a combination rule that is somewhat like MBT, but which is somewhat excessive in respect to MBT.

The hypothesis of a "Best-Guess" tendency in multistage inference is clearly supported by the "unaware" *Ss*. Evidently, perhaps because of the complexity of the situation, some *Ss* tend to concentrate almost exclusively on the most likely event in subsequent stages in inference. The Best Guess effect in multistage inference, like conservatism in single-stage inference (Edwards, 1966), seems to be another example of a general inability to combine complicated information. As much of human information processing is multistaged and probabilistic in nature, it would seem that the next appropriate step for application of Bayes' theorem is to find ways of preventing people from making the mistake of ignoring all but the most likely of the intermediate-level events.

REFERENCES

DODSON, J. D. Simulation system design for a TEAS simulation research facility. AFCRL 1112, PCR R-194, Planning Research Corporation, Los Angeles, California, 1961.

EDWARDS, W. Nonconservative probabilistic information-processing systems. University of Michigan, Institute of Science and Technology Report, 5893-22-F, December, 1966.

GETTYS, C., & WILLKE, T. A. The application of Bayes' theorem when the true data

state is uncertain. *Organizational Behavior and Human Performance,* 1969, **4,** 125–141.

HOWELL, W., GETTYS, C., & MARTIN, D. On the allocation of inference functions in decision systems. *Organizational Behavior and Human Performance,* 1971, **6,** 132–149.

SNAPPER, K., & FRYBACK, D. Inferences based on unreliable reports. *Journal of Experimental Psychology,* 1971, **87,** 401–404.