Bulletin of the Psychonomic Society
1977, Vol. 9 (2), 106-108

Hypothesis testing in stimulus integration tasks of varying difficulty

KENT L. NORMAN
University of Maryland, College Park, Maryland 20742

The processes of hypothesis selection and testing were investigated in a simple multiple-cue learning task. Subjects learned to predict the numeric value of a criterion on the basis of a set of cues. The criterion was computed as the average of two of the cues plus random error. Following each trial, subjects were asked to select the two cues that they thought were relevant. The number of cues and the predictability of the criterion were varied factorially but, within the limits manipulated, did not affect performance to any appreciable degree. The results suggest that a subjective evaluation function based on the average of pairs of cues was operative and that a cut point was employed to decide whether to maintain or reject the current hypothesis.

The processes of hypothesis selection and testing apply to many forms of learning and problem solving tasks. Consider a scientist observing the relationship between a set of variables and a criterion variable that he is trying to predict. Over a number of observations, he may reject some variables as irrelevant and retain others as important. At each observation the scientist entertains a current hypothesis regarding the relationship.

Within such a process the important issues relate to the selection, usage, and testing of possible hypotheses. For example, how does a person select or generate a new hypothesis from past observations? Given that an incorrect hypothesis is being temporarily maintained, how does the individual make the best use of it? Finally, what does it take to reject a hypothesis in favor of another? Bruner, Goodnow, and Austin (1956) and others have investigated such questions within the concept attainment task. The situation outlined above differs from most concept attainment tasks in that the cues and the criterion classification are interval in nature rather than nominal. However, enough similarities exist to suggest that hypothesis testing strategies may be used by subjects in both tasks.

The task employed in the present study is known formally as multiple-cue learning. The subject is presented an array of cues. Over a number of trial observations, he is to learn to predict the value of a criterion. The task bears a striking similarity to the statistical problem of multiple regression, and for this reason correlational models are typically used to describe and assess performance (e.g., Hammond & Summers, 1972).

Multiple-cue learning is an information integration task in which feedback is provided following each judgment. Subjects must integrate or combine the values of cues on each trial. Figure 1 presents a schematic of the general integration process. An array of stimulus information is presented to the subject. Each piece of information attains an internal representation in the form of a subjective value. These values are then integrated to produce a single internal judgment. The integration function is an algebraic composition rule which may take on a form such as a sum or an average. The resultant internal judgment is then mapped to an external response value. To this point, the schematic is the basic outline of Anderson's (1974) information integration theory. What is new is an allowance for the effect of feedback and a decision mechanism that governs changes in the system. Such changes could alter the internal representation of the stimuli, the integration function, or the response function. As a first approximation, this mechanism is assumed to operate in a hypothesis generation and test mode. Studies by Norman (1974a, b) lend support to the hypothesis

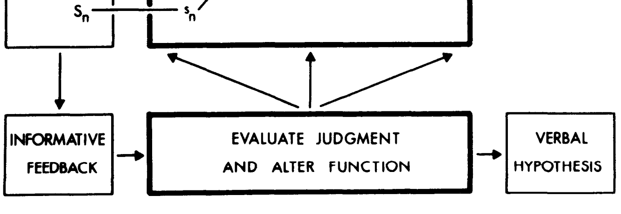

Figure 1. Schematic of an information integration model incorporating a feedback mechanism.

Requests for reprints should be addressed to Kent L. Norman, Department of Psychology, University of Maryland, College Park, Maryland 20742. The author wishes to thank Michael Pruitt for his assistance in conducting this experiment.

testing strategy in a psychophysical averaging task. Subjects learned to change the weighting of stimulus magnitudes in the integration rule as a function of feedback.

The present experiment investigated hypothesis testing in a multiple-cue learning task that required subjects to discover which cues were relevant. The number of irrelevant cues and the predictability of the criterion were varied.

METHOD

On each trial the subject was shown a horizontal array of numbers displayed in windows labeled A, B, C, etc. The subject was told that the average of two of these numbers could be used to predict the criterion number. After the subject made his prediction, he was shown the value of the criterion in a window located below the array of numbers. Following each trial, the subject indicated his current hypothesis as to which positions were relevant. Trials were terminated after the subject named the correct positions five times in succession or exceeded 30 trials. Subjects worked on one problem for practice and then completed three test problems. Two different positions were randomly selected as the relevant cues for each of the problems from the total number of positions presented to the subject. Either two, three, or four of the cues were irrelevant.

The numbers in the array were random integers from 1 to 10, generated according to a uniform probability distribution. The criterion number was the average of the two relevant numbers plus either low or high uniform variability. Under the low error variability condition, the range of random error was 3; under the high error variability condition, it was 5. In addition, the criterion was constrained to an integer value ranging from 1 to 10.

The number of irrelevant positions and the error variability of the criterion number were varied between subjects in a 3 by 2 factorial design. Ten subjects (five male and five female) served in each of the six groups. All subjects volunteered from experimental psychology classes at the University of Alabama and were randomly assigned to conditions in a sequential manner.

RESULTS AND DISCUSSION

Studies in concept attainment show a geometric increase in the difficulty of solving problems as a function of the number of irrelevant cues (Bourne, 1966). This effect was not found. The number of trials to the last error in choosing the relevant positions increased with the number of irrelevant cues. The means for two, three, and four irrelevant cues were 6.38, 8.85, and 10.17, respectively. However, the difference was not significant [$F(2,54) = 2.29$, $.10 < p < .20$]. The lack of a reliable effect here is probably due to the fact that the subject is given much more feedback information than just whether his response is right or wrong. The criterion number may serve to eliminate several hypotheses on any one trial.

The main effect of criterion predictability was also nonsignificant. The mean number of trials to criterion for low and high error variability were 7.81 and 9.12 [$F(1,54) < 1$]. It was expected that high-criterion error variability would decrease performance. However, within the levels employed, no reliable effect was found.

In order to investigate the problem solving process in more detail, prediction responses were analyzed in conjunction with choices of positions on each trial. First, backward learning curves with respect to the trial of the last error were computed and are shown in Figure 2. Absolute error between the prediction and the criterion is high and constant before the trial of the last error. On the trial of the last error in selecting the relevant positions, error was even greater. After solution, error drops to a minimum, due in part to the error variability in the criterion. Increased error prior to solution and the sharp drop after solution suggests that subjects were using a hypothesis testing strategy. Solution is most likely following a large deviation between the response and the criterion. The prediction response can also be compared to the average of the two relevant cues. This curvé shows the same form but eliminates the component of error variability in the criterion.

When the absolute deviation between the average of the positions selected and the response is plotted in Figure 2, there is little or no change over trials. The two values agree closely, almost within round-off error. This indicates that subjects consistently generated their prediction responses by taking the average of the two cues currently hypothesized as relevant.

At this point a rudimentary model of hypothesis selection and testing should be proposed. Suppose that the subject employs a subjective evaluation or goodness-of-fit function. In the present case this would be some monotonic function of the deviation between the average of two cues and the criterion. Hypothesis selection would proceed by a repeated application of the function across a set of possible combinations of positions. The subject then selects the combination leading to the minimum deviation from the criterion. The results of this experiment indicate that when subjects selected a new hypothesis before solution, the combi-

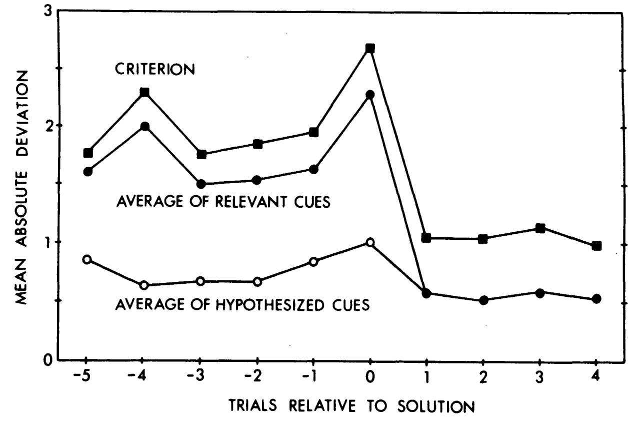

Figure 2. Mean absolute deviation of response from criterion, average of relevant cues, and average of hypothesized cues as a function of trials relative to solution.

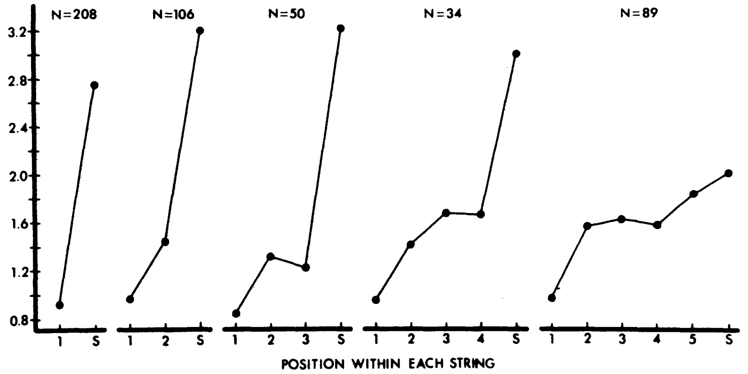

Figure 3. Mean absolute deviation of criterion from the average of the cues selected as relevant as a function of the number of trials on which the hypothesis is retained and the trial on which it is rejected.

nation was generally among the best 15% of all possible combinations. However, when the mean number of better hypotheses are considered instead of the percent, an interesting difference emerges due to the number of irrelevant cues. The mean number of better hypotheses was .68, 1.57, and 2.09 for groups with two, three, and four irrelevant cues [$F(2,54) = 11.60$, $p < .01$]. It is likely that on any one trial subjects test only a limited number of cue combinations and do not exhaust their search for the best combination among the total possible.

A hypothesis is retained as long as it is within a subjective tolerance limit. It is rejected if the deviation between the average of these two cues and the criterion is too large. Figure 3 shows this absolute deviation for runs of the same hypothesis from the initial trial on which the positions were selected to the trial after which the subject eliminated that hypothesis. In each run length the deviation was at a minimum on the trial that the positions were selected. A higher deviation was tolerated on later trials, but when the hypothesis was rejected, the deviation was extreme. It should also be noted that this deviation tended to be higher for runs of two and three than for shorter or longer runs. Subjects were less likely to give up a hypothesis supported on several trials. The form of these results support quite clearly the use of a subjective goodness-of-fit function on the part of the subjects.

One important question is how subjects set their tolerance levels for evaluating hypotheses. Subjects are aware of the usual variability of the cues and the criterion. Studies show that subjects can provide estimates of variability (Peterson & Beach, 1967). It is likely that subjects also develop intuitive values for standard errors and confidence intervals. Subjects, however, did not appear to develop different levels of tolerance under the conditions of low and high error variability in prediction of the criterion. This is shown by the finding that the error in prediction prior to a switch was about equal for the two conditions. It took about the same amount of deviation between the response and the criterion for subjects in both conditions to reject a hypothesis. This led to a slight increase in false rejections of the correct hypothesis in the high-error condition over the low-error condition. Overall, this had little effect on performance since subjects usually resampled the correct hypothesis on later trials.

In conclusion, the findings support a hypothesis testing strategy in multiple-cue learning. Subjects set up an evaluation function that can be used to select and reject hypotheses. Such a strategy appears to optimize performance and in the present case may have nullified the effects of increasing the number of irrelevant cues and error variability of the criterion. In the present task, the set of possible hypotheses concerning the prediction process was limited. Subjects were told that the integration rule was an average and that they had only to learn which two cues were used to compute the average. In more complex multiple-cue learning tasks, subjects may have to consider hypotheses concerning the valuation of cues, the algebraic combination, and the response function. It is suggested that the same type of subjective evaluation function used by subjects in the present task will also be used in more complex tasks.

REFERENCES

Anderson, N. H. Algebraic models in perception. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. 2). New York: Academic Press, 1974.

Bourne, L. E., Jr. *Human conceptual behavior.* Boston: Allyn and Bacon, 1966.

Bruner, J. S., Goodnow, J. J., & Austin, E. A. *A study of thinking.* New York: Wiley, 1956.

Hammond, K. R., & Summers, D. A. Cognitive control. *Psychological Review*, 1972, **79**, 58-67.

Norman, K. L. Dynamic processes in stimulus integration theory: The effects of feedback on the averaging of motor movements. *Journal of Experimental Psychology*, 1974, **102**, 399-408. (a)

Norman, K. L. Rule learning in a stimulus integration task. *Journal of Experimental Psychology*, 1974, **103**, 941-947. (b)

Peterson, C. R., & Beach, L. R. Man as an intuitive statistician. *Psychological Bulletin*, 1967, **68**, 29-46.

(Received for publication September 15, 1976.)