Numerosity in probability learning and decision making with multiple equivalent predictors

BRUCE L. HANSON and LOWELL M. SCHIPPER
The Pennsylvania State University, University Park, Pa. 16802

Multiple probability learning with four predictors and a single event to be predicted used five treatments with πs = .10, .30, .50, .70, or .90 for each of the four predictors, respectively. Subsequent decision making consisted of predictions without feedback given each of the 15 combinations of the four predictors within each treatment. In decision making, significantly more predictions of the event were made to the foursome than to combinations of less than four. Additionally, triples resulted in reliably more predictions than doubles. The "end effect" of the foursome is reminiscent of, but not the same as, that found by others. Additionally, reliable sex differences were found in probability learning for the .10 and .30 treatments.

The primary purpose of this study was to examine the effects of numerosity, per se, of equivalent probabilistic cues used to make decisions.

In earlier studies (Schipper, 1966, 1967) Ss initially learned the probability of some outcome associated with each of several cues, or predictors, in a standard two-alternative, noncontingent probability learning situation. This was followed by decision making which consisted of predicting the occurrence or nonoccurrence of that outcome, or event, without feedback, given various *combinations* of predictors. It appeared that Ss carried out some form of nonlinear averaging of the probabilistic cues in decision making (Schipper, 1967). The chief difficulty with this averaging explanation arose when all of the predictors in the display appeared simultaneously in a decision-making trial. This presentation of the entire array of predictors resulted in a level of responding not predictable by the averaging hypothesis used for combinations of fewer than all predictors. Since the π values are different for each predictor, an exact measure of this effect could not be determined.

Anderson (1967) observed similar behavior in Ss' evaluations of sets of adjectives describing hypothetical personalities. Using a maximum number of six or nine adjectives, he found Ss' evaluations were more extreme when they were given presentations of the entire set of six or nine than when presented with any portion of the set. He called this an end effect.

The same analytic difficulty arises from Anderson's results as from Schipper's. Since each cue (adjective or single predictor light) is different from every other one, the effect of number of predictors is confounded with the effect of their cueing or predictive "values." The use of several predictors of equal π value avoids this difficulty by separating numerosity and cue value effects.

DESIGN

The display consisted of a horizontal array of four green lights mounted directly above a single red light. Ss were assigned to one of the following five treatments: T_1, $\pi = .10$; T_2, $\pi = .30$; T_3, $\pi = .50$; T_4, $\pi = .70$; T_5, $\pi = .90$. π for all four predictors in a single treatment was, of course, the same.

In Session 1, Ss were given 120 probability learning trials on each of the four predictors. Occurrences of the four predictors were completely randomized within the set of 480 trials. Occurrences of the event (red) light were assigned randomly to each consecutive series of 10 presentations of each predictor according to treatment π values. In T_3, the red light came on 5 trials of each 10-trial block. In T_2, 3 of the 5 event trials of T_3 were followed by the red light, and in T_1 the red light occurred on one of those three trials in T_3. T_1 and T_5 were complementary, as were T_2 and T_4, i.e., the event occurred for the one treatment if, and only if, it did not occur for the other.

The second session, the following day, consisted of 200 additional learning trials followed by 150 decision-making trials. The 200 learning trials consisted of 50/predictor, randomized in the same way as before.

Following the 200 learning trials, Ss were presented 150 decision-making trials, 10 for each of the 15 combinations of the four predictors. Presentations of the combinations

were completely randomized within the series of 150 trials. The red light was covered by black tape, so Ss had no way of knowing whether or not the event occurred.

Appearances of the lights were programmed on punched paper tapes read by two Western Union tape readers. Intervals were controlled by four Hunter timers, Model 100-B.

SUBJECTS

Ss were 122 volunteers from the introductory psychology classes at The Pennsylvania State University, 66 males and 56 females, none of whom had previously participated in a decision-making experiment. They were run in 13 groups, ranging in size from 5 to 16. All groups consisted of both males and females, except for one group of eight women in T_2. A treatment was considered closed when at least 10 males and 10 females had provided data for that treatment. Distribution of Ss was: T_1, N = 10 males, 11 females; T_2, N = 14 males, 11 females; T_3, N = 15 males, 11 females; T_4, N = 14 males, 10 females; T_5, N = 13 males, 13 females.

PROCEDURE

On each trial, one (or more, if decision making) of the four green lights appeared for 4 sec. On appropriate trials, the red light came on 3 sec after the onset of the green light, and both stayed on together for 1 sec. Ss made their predictions by marking one of two alternative spaces on answer sheets during the 3-sec interval after a green light appeared but before the red light might have appeared. The intertrial interval was 2 sec.

Before the probability-learning sessions, Ss were told to predict whether the red light would or would not come on and to "try to be correct as often as possible." Before decision making, Ss saw the red light covered and were told that the red light would still either occur or not occur and that they were to predict this outcome, but now they would sometimes see more than one green light. They were instructed to "use the information . . . learned in the first part of the study" and, again, to "try to be correct as often as possible," although they wouldn't know whether they were correct or not.

Table 1
Mean and SD of Percentage Yes Responses for the Last Five Blocks of Probability Learning

| Treat- | Males | | Females | |
ments	Mean	SD	Mean	SD
10%	14.4	14.7	4.0	6.3
30%	23.4	22.7	30.5	19.6
50%	53.4	14.8	52.8	11.1
70%	76.1	14.0	75.0	16.0
90%	92.4	14.6	96.6	6.2

Table 2
Mean Percentage Yes Responses for the Four Numerosity Levels in
Decision Making for Each Treatment and Sex

	Treatment					
	10%	30%	50%	70%	90%	
			Singles			
Males	18	27	53	72	92	52.0
Females	6	38	45	59	94	
			Doubles			
Males	13	20	47	64	92	50.6
Females	7	28	50	78	96	
			Triples			
Males	13	29	54	76	93	55.5
Females	9	36	62	73	96	
			Foursome			
Males	31	46	61	83	92	64.9
Females	22	46	75	87	96	
	11.7	29.4	52.3	71.0	93.9	

Note—Entries to the right of numerosity levels are means for all Ss at each level. Entries at the bottom of each treatment column represent means for both sexes across the four numerosity levels. Each experimental observation was given equal weight in calculating these means. Due to the different numbers of combinations in the numerosity categories, each S contributed 40 responses to singles and triples, 60 to doubles, and 10 to the foursome.

RESULTS

The 480 learning trials on Day 1 were divided into 12 blocks of 40, with each block containing 10 trials with each predictor. Since the presentations of the predictors were randomized within the entire 480-trial sequence, a block need not have contained 40 adjacent trials, e.g., Block No. 3 contained Trials 21-30 for each predictor, but not necessarily Trials 81-120 overall. Each block did, of course, contain the appropriate number of event occurrences. The 200 Day 2 trials were similarly divided into five blocks of 40. Figure 1 shows mean percentage yes responses (predictions that the red light would occur) by block, treatment, and sex. Means and standard deviations of responses summarized over the five Day 2 blocks by treatment and sex are given in Table 1.

The decision-making data for the 15 possible combinations of predictors were grouped according to the four levels of numerosity: singles, the four individual predictors; doubles, the six combinations of two; triples, the four combinations of three; and the single foursome. Mean percentage responses for each numerosity level, by sex and treatment, appear in Table 2.

DISCUSSION

Probability Learning

Analysis of the last five blocks of training showed differences among the five treatments to be highly reliable, $F(4,112) = 474.90$, $p < .001$. Since there were no reliable differences among these five blocks and since mean learning curves in Fig. 1 were fairly stable after eight or nine blocks, it was concluded that sufficient learning had taken place to provide a basis for decision making.

The main effect of sex was not statistically significant, but the Sex by Treatments interaction was highly reliable, $F(4,112) = 4.06$, $p < .005$. This interaction can be seen in Fig. 1 as the reversal of the relative positions of males and females from treatment to treatment. Individual F comparisons, based on Student's range statistic, between the sexes for each treatment showed statistically significant differences for the 10%,

$F(1,112) = 9.18$, $p < .01$, and 30%, $F(1,112) = 5.24$, $p < .05$, treatments only.

Wallach & Kogan (1959), from their own research and a review of the work of others, concluded that females were more conservative than males in decision-making situations with high uncertainty but were less conservative than men when uncertainty was low. If we assume that certainty is proportional to the proximity of values to zero and unity and that conservatism means responding at a neutral, i.e., about 50% rate, then these data are quite compatible with those of Wallach and Kogan. On the other hand, defining conservatism as maximizing correct responses by always choosing the more probable outcome makes these data quite incompatible with their conclusions. Whatever the explanation of these sex differences, they appear to be real and reliable, and generalization of probability learning data from one sex to the other probably is dangerous.

Decision Making

Analysis of the decision-making data showed a significant difference among the probability levels, $F(4,112) = 235.50$, $p < .001$, but an absence of a statistically significant sex main effect or an interaction.

The effects of numerosity also were highly reliable, $F(3,336) = 13.96$, $p < .001$. Individual F comparisons, the same as those discussed above in probability learning, showed the foursome to be different from all other numerosity levels [foursome vs triples, whose mean was nearest that
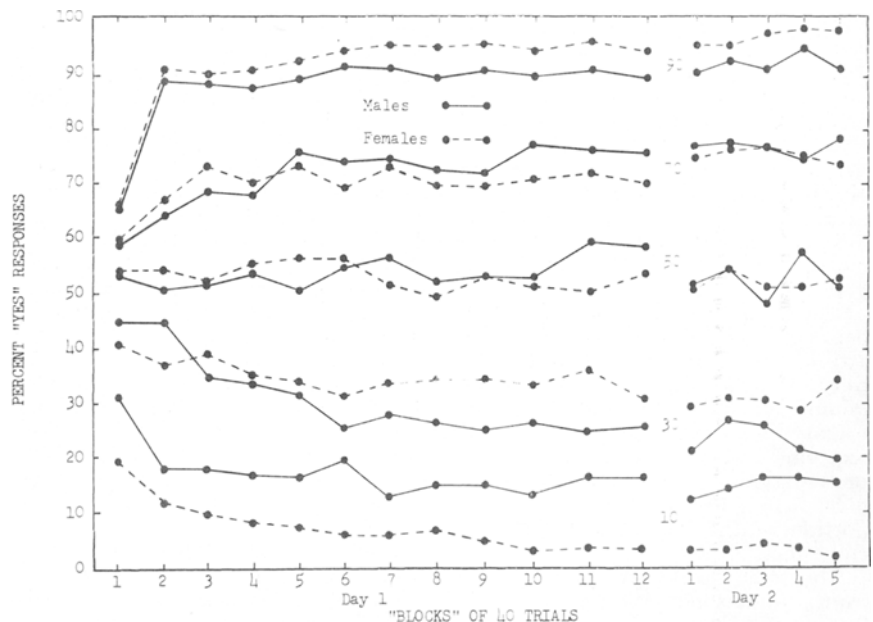

Fig. 1. Mean percentage yes responses for each block by sex and treatment.

of the foursome: F(1,336) = 14.74, p < .005]. Additionally, the doubles and triples were different from one another, F(1,336) = 3.16, p < .05.

Thus, an end effect was shown, but not quite the same as Anderson's. His Ss' ratings of likability of the hypothetical personality became *more extreme* when the entire set of six or nine adjectives was presented. Yet Ss in this study predicted the occurrence of the event light more often with all predictors present with all treatments. For $\pi < .50$ this is less extreme rather than more extreme behavior.

Anderson's task seems, intuitively, to be less ambiguous than the one in this study. A reasonable explanation for his end effect may be that Ss felt they were given all the information available about the personality when an entire adjective set was presented and could have felt more certain whether they liked or disliked the personality. Presentation of only a subset may have implied there was additional information on which the judgment should have been based, but the information was unavailable. In contrast, it seems there is no a priori reason to suppose that a greater

number of predictor lights in this study provided Ss with greater surety that the event would occur.

The general numerosity effect seems to be of an unspecified method of averaging, as Schipper has suggested. The difference between the doubles and triples is not immediately explainable. Anderson suggests that Ss average the adjectives with some subjective impression (possibly analogous to some probabilistic response set), but this predicts that judgments would be a monotone increasing (or decreasing) function of the number of cues, which, in this study, is not the case.

REFERENCES
ANDERSON, N. H. Averaging model analysis of set-size effect in impression formation. Journal of Experimental Psychology, 1967, 75, 158-165.
SCHIPPER, L. M. Context effects in probability learning and decision-making. Psychological Reports, 1966, 18, 131-138.
SCHIPPER, L. M. Extreme probabilities in learning and decision making. Journal of Experimental Psychology, 1967, 73, 149-151.
WALLACH, M. A., & KOGAN, N. Sex differences and judgment processes. Journal of Personality, 1959, 27, 555-564.

Redundant stimulus coding and keeping-track performance

DAVID G. ALDEN, JACELYN R. WEDELL, and ARNOLD F. KANARICK
Systems and Research Center, Honeywell, Inc., Minneapolis, Minn. 55113

The effect of redundant color coding on keeping-track performance was investigated, using symbols as the primary cue. Ss were assigned to one of four coding conditions: symbol color (SC), color symbol (CS), symbol only (S), or color only (C). Performance was measured on a task requiring S to monitor eight information channels, which were being updated randomly. Results indicated that the addition of a redundant color code did not yield a significant improvement in performance, as compared to performance on the component codes (i.e., the S and C groups).

A keeping-track task is one in which S is required to remember the present state of each of a number of variables. The task is a continuous one in which the states of the variables are changed at random intervals. This experimental paradigm was introduced by Yntema & Mueser (1960) and has the advantage of enabling one to study a dynamic or a continuous memory process.

Kanarick & Petersen (in press) investigated the effect of redundant color coding, in which the stimulus was uniquely identified by both a primary dimension and a secondary color code, on keeping-track performance. They hypothesized that the use of color as a redundant cue would increase the effectiveness of keeping-track performance. Their

study required S to keep track of information consisting of numbers, colors, or numbers on a colored background, the latter being the redundantly coded information. Their results did not support the hypothesis. Rather, the technique appeared to increase performance variability through the number of strategies Ss used. The authors suggested that the apparent inefficacy of the redundant cue may be a result of the ease with which numbers can be encoded and chunked. The present study was, therefore, an attempt to investigate the effect of redundant color coding on keeping-track performance, using additional task controls.

Symbols were chosen for one stimulus dimension because they are of a class that does not possess a

natural order, as compared with numbers. Monty, Fisher, & Karsh (1967) suggest that use of a class of stimuli possessing sequential order (such as numbers or letters of the alphabet) as a coding dimension is likely to lead to greater proficiency in keeping-track performance than use of stimuli not possessing such an order.

To reduce variability of individual keeping-track strategies, paced overt rehearsal during update trials was introduced.

METHOD
The Ss were male and female students from the University of Minnesota and Hamline University. All Ss were pretested for color defectiveness by the Farnsworth D-15 test; two Ss were eliminated by this test. The remaining Ss were assigned randomly to one of four coding conditions, with eight Ss per group. The four coding conditions were: symbol color (SC), symbol being the cue S was instructed to attend to and color serving as the redundant cue; color symbol (CS), the cues being reversed; symbol only (S); and color only (C).

The apparatus consisted of eight IEE solid-state digital readouts controlled by a tape reader and interval timers. The readouts, each 2 in. square, were mounted horizontally on a rack approximately 40 in. from S's eye level. Each readout was programmed to display either a symbol (=, >, +, —, ÷, or ×), a color (yellow, blue, purple, green, red, or orange), or a unique symbol-color combination (=/yellow, >/blue, +/purple, —/green, ÷/red, or ×/orange). Small white lights to pace rehearsal and 1-in. capital identification letters (A through H) were mounted above the readouts or "channels."

Each S was given paired-associate training with the six unique combinations of colors and symbols to the criterion of two errorless trials. All Ss received equal training regardless of coding condition. The S was then told that his task would be to keep track of many pieces of information at one time. One new piece of information would be presented to him in one of the windows (readouts) at a time. The S was instructed that the information would be in the appropriate mode (i.e., symbol and color, color only, or symbol only) for his group. Following each presentation of new information in one of the channels (update trial), he would be asked to recall what information had most recently been stored in a channel (interrogation trial). These two types of trials would alternate.

The new updated information remained displayed throughout each