THEORETICAL NOTE: ALL-OR-NONE LEARNING AND INTERTRIAL FORGETTING

by

Richard C. Atkinson and Edward J. Crothers

TECHNICAL REPORT NO. 58

July 24, 1963

PSYCHOLOGY SERIES

INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES

Applied Mathematics and Statistics Laboratories

STANFORD UNIVERSITY

Stanford, California

THEORETICAL NOTE: ALL-OR-NONE LEARNING AND INTERTRIAL FORGETTING[1]

Richard C. Atkinson and Edward J. Crothers

Stanford University

Abstract

Several alternative interpretations of all-or-none processes for paired-associate learning and concept formation are examined. These models, along with a simple incremental process, are compared to data obtained in three-response and four-response paired-associate learning experiments. The results strongly favor a two-process model that postulates a distinction between long-term and short-term retention and assumes an intertrial forgetting process.

In recent issues of this journal Estes (1960), Bower (1962), and Suppes and Ginsberg (1963) have examined a wide array of data on paired-associate learning and concept formation in terms of an all-or-none process. The particular model they consider represents a special case of more general models of Stimulus Sampling Theory, and has been frequently labeled as the one-element pattern model. In a paired-associate experiment the single stimulus element represents a stimulus item from a list of paired-associates; in a concept formation experiment the stimulus element represents a concept, or some aspect of a concept. The two assumptions of the model are as follows: (1) Until the stimulus element is conditioned, there is a constant probability g that the subject will respond correctly; (2) On each trial there is a probability c that the single element will become conditioned to the correct response. Thus, on trial n of an experiment the stimulus element can be regarded as being in one of two conditioning states: In state C the element is conditioned to the correct response; in state $\overline{C}$ the element is unconditioned. The element starts out in state $\overline{C}$ and moves between these two states as specified by the transition matrix

$$\begin{array}{c} \\ C \\ \overline{C} \end{array} \begin{array}{c} \begin{array}{cc} C & \overline{C} \end{array} \\ \begin{bmatrix} 1 & 0 \\ c & 1-c \end{bmatrix} \end{array} . \qquad [1]$$

By and large, the results reported by Estes, Bower, and Suppes and Ginsberg indicate a remarkably close correspondence between observed and predicted values for the one-element model. The correspondence is particularly impressive when compared to goodness-of-fit results obtained

for other models. However, despite the excellent fits of the one-element model, there is at least one aspect of the data that is contradictory. As pointed out by Suppes and Ginsberg, when appropriate statistical analyses are made one can often demonstrate a non-stationary effect before the last error; i.e., there is a tendency for the probability of a correct response to increase over trials prior to the last error and not simply remain a constant g, as predicted by the theory.

To account for this non-stationary effect, Suppes and Ginsberg propose a two-element stimulus sampling model. Roughly speaking, their model is defined by three conditioning states: C_0, C_1, and C_2. For state C_0 both elements are unconditioned and the probability of a correct response is g; for state C_1 one of the two elements is conditioned and the probability of a correct response is g'; for state C_2 both elements are conditioned and the probability of a correct response is 1. Applying stimulus-sampling axioms, they derive the transition matrix

$$\begin{array}{c} \\ C_2 \\ C_1 \\ C_0 \end{array} \begin{array}{c} \begin{array}{ccc} C_2 & C_1 & C_0 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ b & 1-b & 0 \\ 0 & a & 1-a \end{bmatrix} \end{array}, \qquad [2]$$

and show that the probability of a correct response over trials before the last error is an increasing function bounded between g and g'. In their view this two-element process represents a conceptual compromise between incremental and all-or-none learning models.

The purpose of this note is to examine an alternative approach to the one proposed by Suppes and Ginsberg that is in the spirit of a one-element analysis, but predicts the non-stationary effect. Because of the particular data to be considered here, the model will be formulated for a task involving a fixed set of r response alternatives; however, generalization of the model to unrestricted response sets presents no new problems. Specifically we shall consider a paired-associate task in which the subject is told the responses available to him; each response occurs equally often as the to-be-learned response, and consequently the probability of a correct response by guessing is $\frac{1}{r}$. On each trial the stimuli are exhibited singly in a new random order. To the presentation of each stimulus the subject is required to make a response and is then informed of the correct response. The experimental procedure is precisely that described by Bower (1961).

Within the framework of an all-or-none model one can introduce the distinction between long-term retention (C_L) and short-term retention (C_S). That is, at any point in time a given stimulus item is viewed as being in one of three states of conditioning. In the unconditioned state $(\bar{C})$, the subject responds at chance, whereas in either conditioned state C_L or C_S he makes the correct response with probability 1. The distinction between the two conditioned states is in terms of a forgetting process. Once an item has passed into state C_L it remains there; but an item in the short-term state can move back into the unconditioned state as a result of forgetting that can occur from one presentation of the item to the next.

Crothers (1963) proposed the notion of a three-state Markov model

with a forgetting parameter and investigated its statistical properties. The general concept of a strong state and weak state of conditioning (which are represented here as long-term and short-term retention states, respectively) and a response probability of unity in either conditioned state has been advanced and tested in a probability learning situation (Atkinson, 1962; Myers and Atkinson, 1964).

The assumptions that we shall consider in this paper governing transitions among the conditioning states are as follows: If an item is presented that is in state $\overline{C}$, then with probability α it goes into state C_L and with probability $1 - \alpha$ it goes into state C_S ; if an item is presented that is already in state C_S , then with probability β it goes into state C_L and with probability $1 - \beta$ it remains in C_S . Thus after each presentation the item is in either state C_L or C_S , and if the item were to be presented again immediately the subject would make the correct response with probability 1 . However, in the typical paired-associate experiment other events intervene from one presentation of an item to its next presentation and during this period we assume there is probability δ than an item in the short-term state will return to state $\overline{C}$. Thus, from one presentation of the item to the next the process can be described by the following transition matrix:

$$\begin{array}{c} \\ C_L \\ C_S \\ \overline{C} \end{array} \begin{array}{c} \begin{array}{ccc} C_L & C_S & \overline{C} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ \beta & (1-\beta)(1-\delta) & (1-\beta)\delta \\ \alpha & (1-\alpha)(1-\delta) & (1-\alpha)\delta \end{bmatrix} \end{array} . \quad [3]$$

For this process, which we shall call the long-short model, it can be shown that the probability of a correct response over trials before the last error is not constant but a function starting at $\frac{1}{r}$ and bounded below 1 .

To evaluate the long-short model (LS-model) against the Suppes and Ginsberg model (SG-model) and the simple one-element model (S-model), we shall examine data from three experiments; one with r equal to 3 and two with r equal to 4 . In Experiment I 65 college students were run using a list of 12 Greek letters as stimuli and the numbers 4, 6, and 8 as responses; in Experiment II 20 college students were run using 16 consonant trigrams as stimuli and the numbers 4, 5, 6, and 7 as responses; in Experiment III 20 grade-school children (sixth grade) were run using a list of 12 highly dissimilar designs as stimuli and the letters A, B, C, and D as responses. Tables 1, 2, and 3 present the observed frequencies for the various possible outcomes over trials 2, 3, and 4 of the experiments; we let e_n and c_n denote an error and a correct response, respectively, on trial n . For the 65 subjects in the first experiment there were $12 \times 65 = 780$ item-response sequences and, as indicated in Table 1, for 317 sequences no errors occurred on trials 2, 3, and 4; 31 displayed no errors on trials 2 and 3 but an error on trial 4, and so forth. Similarly for the 20 subjects in the second experiment there were 320 item-response sequences, and for the 20 subjects in the third experiment there were 240 item-response sequences.

For the LS-model described by Equation 3 we can derive expressions for the event sequences displayed in Tables 1, 2 and 3. Namely,

Table 1

Observed and Predicted Outcome Frequencies

for Experiment I

Outcome Sequences	Observed Frequencies	Predicted Frequencies			
		LS-model	SG-model	S-model	L-model
$c_2c_3c_4$	317	315.1	300.5	301.9	226.7
$c_2c_3e_4$	31	38.5	35.2	20.8	62.7
$c_2e_3c_4$	71	65.5	59.9	46.1	104.3
$c_2e_3e_4$	50	44.9	32.9	41.5	28.8
$e_2c_3c_4$	140	142.7	161.1	152.9	191.8
$e_2c_3e_4$	42	44.9	40.5	41.5	53.0
$e_2e_3c_4$	80	76.2	86.5	92.2	88.2
$e_2e_3e_4$	49	52.2	63.3	83.0	24.4
minimum χ^2		3.15	18.86	37.67	120.01

Table 2

Observed and Predicted Outcome Frequencies

for Experiment II

Outcome Sequences	Observed Frequencies	Predicted Frequencies			
		LS-model	SG-model	S-model	L-model
$c_2c_3c_4$	120	119.8	113.9	116.4	75.7
$c_2c_3e_4$	11	9.5	11.3	5.2	24.4
$c_2e_3c_4$	18	19.9	19.9	14.0	41.6
$c_2e_3e_4$	15	18.2	13.0	15.6	13.4
$e_2c_3c_4$	61	61.7	67.3	64.1	80.6
$e_2c_3e_4$	19	18.2	16.3	15.6	26.0
$e_2e_3c_4$	42	38.1	40.6	42.1	44.2
$e_2e_3e_4$	34	34.7	37.6	46.9	14.3
minimum χ^2		1.44	2.27	12.12	80.94

Table 3

Observed and Predicted Outcome Frequencies

for Experiment III

Outcome Sequences	Observed Frequencies	Predicted Frequencies			
		LS-model	SG-model	S-model	L-model
$c_2c_3c_4$	81	80.5	69.8	68.7	47.0
$c_2c_3e_4$	16	12.7	13.3	5.2	19.0
$c_2e_3c_4$	15	18.0	18.9	11.3	30.8
$c_2e_3e_4$	12	19.3	11.3	15.6	12.4
$e_2c_3c_4$	36	33.7	48.6	42.8	56.3
$e_2c_3e_4$	22	19.3	15.8	15.6	22.7
$e_2e_3c_4$	27	27.3	32.2	33.9	36.9
$e_2e_3e_4$	31	29.2	30.2	46.8	14.9
minimum x^2		4.76	9.75	37.13	60.58

$$\Pr(c_2c_3c_4) = \alpha + (1-\alpha)(1-\delta)[\beta + (1-\beta)(1-\delta)(1-Y+Yg) + Yg(1-Z+Zg)]$$
$$+ Zg[\alpha + (1-\alpha)(1-\delta)(1-Y+Yg) + Zg(1-Z+Zg)]$$
$$\Pr(c_2c_3e_4) = (1-\alpha)(1-\delta)(1-g)[(1-\beta)(1-\delta)Y + YZg] + Zg(1-g)[(1-\alpha)(1-\delta)Y + Z^2g]$$
$$\Pr(c_2e_3c_4) = (1-g)(1-Z+Zg)[(1-\alpha)(1-\delta)Y + Z^2g]$$
$$\Pr(c_2e_3e_4) = (1-g)^2Z[(1-\alpha)(1-\delta)Y + Z^2g]$$
$$\Pr(e_2c_3c_4) = (1-g)Z[\alpha + (1-\alpha)(1-\delta)(1-Y+Yg) + Zg(1-Z+Zg)]$$
$$\Pr(e_2c_3e_4) = (1-g)^2Z[(1-\alpha)(1-\delta)Y + Z^2g]$$
$$\Pr(e_2e_3c_4) = (1-g)^2Z^2(1-Z+Zg)$$
$$\Pr(e_2e_3e_4) = Z^3(1-g)^3 \qquad [4]$$

where, to simplify the expressions, we have introduced the following notation: $Y = (1-\beta)\delta$, $Z = (1-\alpha)\delta$, and $g = \frac{1}{r}$. See Atkinson and Estes (1963) for a discussion of the methods employed in these derivations.

It is important to realize that the LS-model reduces to the one-element model described by Equation 1 when $\delta = 1$. Hence to obtain appropriate expression for the S-model, simply set $\delta = 1$ and $\alpha = c$ in Equation 4. Expressions for the SG-model can not be obtained by direct substitution in Equation 4, but they are easy to derive and will not be presented here.

In order to make predictions for the data displayed in Tables 1, 2 and 3 we need estimates of the parameters α , β , and δ for the LS-model, g' , a and b for the SG-model, and c for the S-model. There are many ways of making these estimates, but for the present problem a simple method, which yields asymptotically efficient and consistent estimates, is to minimize χ^2 . To illustrate the method, let $p_i(\alpha,\beta,\delta)$

denote the theoretical expressions for the response probabilities given in Equation 4, where the subscript i refers to the event listed in row i of Tables 1, 2, and 3. Further, let O_i denote the observed frequencies in row i of a table, and let $T = O_1 + O_2 + \ldots + O_8$. Then we define the function

$$\chi^2(\alpha,\beta,\delta) = \sum_{i=1}^{8} \frac{[Tp_i(\alpha,\beta,\delta) - O_i]^2}{Tp_i(\alpha,\beta,\delta)} \quad , \quad [5]$$

and select our estimates of α, β and δ so that they jointly minimize the χ^2 function. A number of problems are involved in carrying out the minimization analytically, and consequently we have programmed a high-speed computer to systematically scan grids of possible parameter values until estimates are obtained that are accurate to three decimal places. Under the null hypothesis it can be shown that this minimum χ^2 has the usual limiting distribution with 4 degrees of freedom.[2] If k parameters are estimated then there are $8 - k - 1$ degrees of freedom (for a discussion of these techniques see Atkinson and Calfee, 1963).

Applying this method of estimation to the LS-model, we obtain χ^2's of 3.16 (4df), 1.44 (4df) and 4.76 (df) for Experiments I, II, and III, respectively. For the first study, the estimates are $\alpha = .242$, $\beta = .250$ and $\delta = .805$; for the second study $\alpha = \beta = .273$ and $\delta = .875$; and for the third study $\alpha = .063$, $\beta = .500$, and $\delta = .734$. The small values of χ^2 are reflected in the excellent fits displayed in Tables 1, 2, and 3 between observed frequencies and the predictions for the LS-model.

Minimizing Equation 5 for the SG-model yields χ^2 values of 18.86 (4df) , 2.27 (4df) , and 9.75 (4df) , for Experiments I, II, and III, respectively. For the first experiment $g' = .922$, $a = .367$ and $b = .094$; for the second study $g' = .930$, $a = .359$, and $b = .070$; and for the third study $g' = .867$, $a = .359$, and $b = .008$. For the S-model the minimum χ^2's are 37.67 (6df) , 12.12 (6df) , and 37.13 (6df) and the corresponding estimates of c are .289 , .297 , and .227 . Predictions for these models based on the above parameter estimates also are displayed in the tables.

Finally, as a basis of comparison we generated predictions for a linear model (see Atkinson and Estes, 1963). The axioms for the linear model (L-model) are quite simple: If we let p_n denote the probability of a correct response on trial n , then $p_{n+1} = (1-\theta)p_n + \theta$ and $p_1 = \frac{1}{r}$. For this model the estimates of θ are .312, .312, and .273 for Experiments I, II, and III, respectively. As evidenced by the poor fits displayed in Tables 1, 2, and 3, the minimum χ^2's are 120.01 (6df) , 80.94 (6df) , and 60.58 (6df) .

The minimum χ^2 values summed over the three experiments provide a measure of the overall fit of the four models. For the LS-model the total χ^2 is 9.35 and is based on 12 degrees of freedom; for the SG-model we have a total χ^2 of 30.88 also based on 12 degrees of freedom; for the S-model, a total of 86.92 (18df) ; and for the L-model a total of 261.53 (18df) . Of course, for all models except the LS-model the total χ^2 values are highly significant. What is particularly impressive in our opinion is that the LS-model yields a total χ^2 value less than

one-third of the value obtained for the SG-model. We should note at this point that the same analyses have been carried out on the 32 possible event sequences for trials 1 through 5 of these experiments and the results agree in detail with the ones reported above.

Admittedly the present analyses are limited; in subsequent research attention will need to be given to other predictions of the LS-model such as the distributions of total errors, number of trials to the first success, trial of the last error, and so forth. However, in view of the present results we feel that a two-process model incorporating an all-or-none conditioning function and intertrial forgetting is worth pursuing. As a first step it will be important to investigate experimentally the effect of such variables as list length, intertrial interval, and stimulus similarity on the estimate of the forgetting parameter δ.[3] Finally, a particularly promising aspect of the LS-model is that it provides a conceptual link between paired-associate learning and current research on memory processes, thus promising a theoretical rapprochement between these two areas of experimentation.

REFERENCES

Atkinson, R. C. Choice behavior and monetary payoff: strong and weak conditioning. In J. H. Criswell, H. Solomon, and P. Suppes (Eds.), Mathematical methods in small group processes. Stanford: Stanford Univer. Press, 1962, 23-34.

Atkinson, R. C., and Calfee, R. C. Mathematical learning theory. In B. B. Wolman and E. Nagel (Eds.), Psychology and the theory of science. New York: Basic Books, 1963, in press.

Atkinson, R. C., and Estes, W. K. Stimulus sampling theory. In R. D. Luce, R. R. Bush and E. Galanter (Eds.), Handbook of mathematical psychology. Vol. 2. New York: Wiley, 1963, 121-268.

Bower, G. H. Application of a model to paired-associate learning. Psychometrika, 1961, 26, 255-280.

Bower, G. H. A model for response and training variables in paired-associate learning. Psychol. Rev., 1962, 69, 34-53.

Crothers, E. J. Markov models for learning with inter-trial forgetting. Technical Report No. 53, Institute for Mathematical Studies in the Social Sciences, Stanford Univer., 1963.

Estes, W. K. Learning theory and the new mental chemistry. Psychol. Rev., 1960, 67, 207-223.

Myers, J. L. and Atkinson, R. C. Choice behavior and reward structure. J. Math. Psychol., 1964, in press.

Suppes, P., and Ginsberg, Rose. A fundamental property of all-or-none models. Psychol. Rev., 1963, 70, 139-161.

FOOTNOTES

1. This article was written while the first author was a Visiting Professor at the University of Michigan and was in part supported by the National Institute of Mental Health (Grant NH-05184) and by the National Science Foundation (Grant 24264).

2. To be more exact, the statistic is χ^2-distributed if we assume that all subjects in a given experiment have the same parameter values, and the item-response sequences are independent. These assumptions are in part justified by the fact that the subjects are drawn from a fairly homogeneous source, and the stimuli are selected so that they are not easily confused.

3. We have applied the LS-model to some data from an experiment in which kindergarten children learned a different list of paired associates each day for five consecutive days. The effect of practice on equivalent lists was reflected in the parameter estimates: α remained relatively constant from day to day, β showed an increase, and δ a marked decrease over days.