PSYCHOMETRIKA—VOL. 22, NO. 1
MARCH, 1957

A STOCHASTIC MODEL FOR ROTE SERIAL LEARNING

RICHARD C. ATKINSON*

INDIANA UNIVERSITY†

A model for the acquisition of responses in an anticipatory rote serial learning situation is presented. The model is developed in detail for the case of a long intertrial interval and employed to fit data where the list length is varied from 8 to 18 words. Application of the model to the case of a short intertrial interval is considered; some predictions are derived and checked against experimental data.

This paper represents a preliminary attempt at quantitative theorizing in the area of rote serial learning. The model is applicable to experimental situations employing the anticipation method [6] and deals with the acquisition of correct responses, anticipatory responses, perseverative responses, and failures-to-respond. In addition, direct applicability of the model is limited to situations restricted as follows: (*a*) moderate presentation rate, (*b*) dissimilar intralist words, (*c*) familiar and easily pronounced words. The explanation for these restrictions is considered later.

Model

The model makes use of the conceptual formulation of the stimulating situation introduced by Estes [3] and elaborated by Estes and Burke [4]. The general assumptions are: (*a*) the effect of a stimulating situation upon an organism is made up of many component events; (*b*) when a situation is repeated over a series of trials, any one of these component stimulating events may occur on some trials and fail to occur on others. Rather than review the rationale of these assumptions, the reader is referred to the Estes-Burke paper which is helpful to an understanding of the present work.

Figure 1 schematically presents the rote serial learning situation. The successive word exposures in a list of $r + 1$ words are indicated by $W_1 , W_2 , \cdots , W_r , W_{r+1}$ where W_1 is the cue for S's first anticipation on each run through the list. R_i' represents a hypothesized covert response associated with the $i + 1$st word presentation; the response of "reading" W_{i+1}. On the other hand, $R_i(i)$ is the response recorded by the experimenter to the ith word presentation and can be either (*a*) a correct anticipation

*The author wishes to thank Professors C. J. Burke and W. K. Estes for advice and assistance in carrying out this research.

†Now at Stanford University.

of the i + 1st word when $j = i$, (b) an incorrect anticipation when $j \neq i$, or (c) a failure-to-respond when the j subscript is omitted. (Symbols and their meanings are listed in Appendix B.)

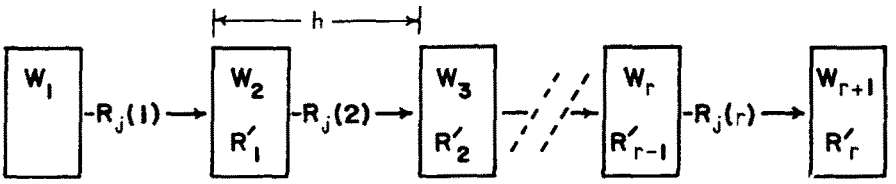

FIGURE 1

Schematic representation of the anticipatory rote serial learning situation.

A period h is defined as the time of a single word exposure, and a trial refers to one run through the list. Since the removal of one word is followed immediately by the presentation of the next, a trial is of time $h(r + 1)$. The intertrial interval is represented as a series of k subintervals each of length h; thus, the intertrial interval is of time kh. When there are $r + 1$ words in a list, the list length is designated as r; this reflects the fact that the r + 1st word is not a cue for an anticipatory response.

The ith word presentation is represented conceptually as a set of stimulus elements S_i, where the sets are pairwise disjoint, and hence the intersection of the $r + 1$ sets is the null set. The number of elements in S_i is N, where N is invariant over i, and a parent set S^* is defined such that the union of the $r + 1$ sets is a subset of S^*. On a given presentation of the ith word a sample of elements from S_i is effective; the likelihood of any element from S_i being in the sample is θ_i where $0 \leq \theta_i \leq 1$. (Derivations presented in this paper are carried out under the simplifying assumption that all elements in S_i are equally likely to occur on any trial.) Therefore, given the ith word presentation, a sample is drawn from S_i of size $N\theta_i$.

Conditional relations, or connections, between response classes and stimulus elements are defined as in other papers on statistical learning theory. The response classes R_1 , R_2 , $\cdots$, R_r , and $\bar{R}$ (failure-to-respond) define a partition of S^* into subsets $S^*_{R_1}$, $S^*_{R_2}$, $\cdots$, $S^*_{\bar{R}}$. Elements in $S^*_{R_i}$ are said to be conditioned to the response class R_i etc. The concept of a partition implies that every element of S^* must be conditioned to either R_1 , R_2 , $\cdots$, or $\bar{R}$, but that no element may be conditioned to more than one. For each element in S_i a quantity $F(i; j; n)$ is defined which represents the probability that an element from set S_i is conditioned to response class R_j at the start of trial n. At times this notation is unnecessarily detailed; the abbreviation $C(i; n)$ is introduced to designate the probability that an element from S_i is conditioned at the start of trial n to a correct anticipatory response.

The anticipatory response at position i on trial n is assumed to be a function of the stimulus elements sampled from S_i on that trial. Specifically, the probability of $R_i(i)$ is the ratio of the number of sampled elements from S_i conditioned to the response class R_i to the number of elements sampled from S_i . Since θ_i is constant for all elements in S_i , the probability of $R_i(i)$ on trial n is the expected value of $F(i; j; n)$.

For each element sampled from S_i on trial n it is postulated that there is: (*a*) a probability λ that the element is returned to S* during the h-interval immediately following the one in which it was sampled; (*b*) a probability $\lambda(1 - \lambda)$ that it it is returned to S* during the second h-interval following the one in which it was sampled; (*c*) a probability $\lambda(1 - \lambda)^2$ that it is returned to S* during the third h-interval following the one in which it was sampled; and so on. The probability that an element will be eventually returned to S* is unity since

$$\sum_{r=0}^{\infty} \lambda(1 - \lambda)^r = 1. \tag{1}$$

The phrase "available at position i" is used to refer to an element sampled from some set and not yet returned to S* during the h-interval in which W_i is presented. The notion of an element being available at a position other than the one at which it was sampled is one way of formalizing the concept of trace stimuli. Parenthetically, note that the probability of an anticipatory response at position i is defined in terms of the stimulus elements sampled from S_i and is not affected by elements which are available at position i but sampled from a stimulus set other than S_i .

The conditioned status of elements sampled from S_i upon their return to S* depends on the anticipatory response made at position i. If a sample is drawn from S_i which elicits a correct anticipatory response, $R_i(i)$, then all elements in the sample become conditioned to the response class R_i and, independent of the time that an element is available, are returned to S* conditioned to that response class. On the other hand, if the sample elicits a response, other than a correct one, all elements in the sample revert to being conditioned to the response class $\bar{R}$, and there is a specified probability that the elements will be conditioned to the R_i' responses which occur before they are returned to S*. That is, given an incorrect anticipation or a failure-to-respond, all sampled elements become conditioned to the response class $\bar{R}$ and then: (*a*) a proportion β of the sampled elements are conditioned to the response class R_i when R_i' occurs, and $(1 - \beta)$ remain unchanged; (*b*) λ of the elements are then returned to S* and $(1 - \lambda)$ remain available during the next h-interval where, again, β of the remaining elements are conditioned to the response class R_{i+1} when R_{i+1}' occurs, and $(1 - \beta)$ remain as they were in the previous interval; (*c*) $\lambda(1 - \lambda)$ are now returned to S* and $(1 - \lambda)^2$ are carried on where β are connected to the response class R_{i+2}

when R'_{i+2} occurs and $(1 - \beta)$ remain as they were in the previous interval; and so on.

Finally, it is assumed that nothing which occurs during the intertrial interval will change the conditional status of the elements not yet returned to S* at the beginning of this interval. That is, elements returned during h-intervals of the intertrial interval have the same conditional status as elements returned in the last h-interval of the list presentation.

More generally stated, if a sample of elements elicits a response which is confirmed as correct (reinforced), then each element in the sample becomes conditioned to that response and will remain conditioned unless the element is sampled at some later trial, and this new sample elicits an incorrect response. If a sample leads to an incorrect response, then the elements in the sample revert to being conditioned to the response class $\bar{R}$ and have a probability β of being conditioned to the response class R_i associated with the R'_i responses which occur before the element is returned to S*. The conditioning proportion β can be interpreted as the probable occurrence of the implicit response R'_i to the i + 1st word presentation. This interpretation does not affect the quantitative formulation of the model.

The present analysis of serial responding requires a modification of the notion of a sampling constant introduced in other papers on statistical learning theory. θ_i is postulated to be a function of the number and order of the words that have preceded the ith word. Once again, consider intervals of time h. If the word exposure has been preceded by an infinite number of h-intervals which do not contain word exposures, then the sampling constant is θ_1 ; if, on the other hand, the word exposure has been preceded by an infinite number of h-intervals each of which contained a word exposure, the sampling constant is θ_∞ . Let $c = \theta_1 - \theta_\infty$, where $c \geq 0$ and, necessarily, $c \leq 1$. Further, designate a decay constant η such that $0 \leq \eta \leq 1$. If a series of successive word exposures occur, and are preceded by an infinite number of h-intervals which do not contain word exposures, then (a) the sampling constant associated with the second word exposure is $\theta_1 - c\eta$; (b) the sampling constant associated with the third word is $\theta_1 - c[\eta + \eta(1 - \eta)]$; ($c$) the sampling constant for the fourth word is $\theta_1 - c[\eta + \eta(1 - \eta) + \eta(1 - \eta)^2]$; and so on. Thus, if the intertrial interval is infinite (i.e., each run through the list is preceded by an infinite number of h-intervals which do not contain word exposures), the sampling constant associated with set S_i on any run through the list is

$$\theta_i = \theta_1 - c[1 - (1 - \eta)^{i-1}]. \tag{2}$$

An inspection of this equation indicates that θ, defined over list positions, has a maximum at position one and approaches $0 \leq \theta_1 - c \leq 1$ as i becomes large.

The formulation of the sampling constant requires a uniform activity

during intervals which do not contain word exposures; θ_i is postulated to be a function of the type of activity.

The equations specified by the above assumptions can now be written. Consider the case in which the intertrial interval is "long," for purposes of the model infinite. This case proves to be simpler than that in which the intertrial interval is "short" because in the infinite interval all elements sampled from S_i on trial n are returned to S* before the beginning of trial $n + 1$ (see equation 1). (Perseverative errors are not possible for the infinite intertrial interval, and their consideration is deferred until discussion of the short interval case.)

Given a list length r and an infinite intertrial interval, the expected values of the probabilities of correct anticipatory responses on trial $n + 1$ to the exposure of W_r , W_{r-1} , and W_{r-2} are

$$C(r; n+1) = (1 - \theta_r)C(r; n) + \theta_r\{C(r; n) + [1 - C(r; n)]\beta\}, \tag{3}$$

$$\begin{aligned} C(r-1; n+1) &= (1 - \theta_{r-1})C(r-1; n) \\ &\quad + \theta_{r-1}\{C(r-1; n) + [1 - C(r-1; n)][\lambda\beta + (1-\lambda)\beta(1-\beta)]\}, \end{aligned} \tag{4}$$

$$\begin{aligned} C(r-2; n+1) &= (1 - \theta_{r-2})C(r-2; n) + \theta_{r-2}\{C(r-2; n) \\ &\quad + [1 - C(r-2; n)][\lambda\beta + \lambda(1-\lambda)\beta(1-\beta) + (1-\lambda)^2\beta(1-\beta)^2]\}. \end{aligned} \tag{5}$$

More generally,

$$C(i; n+1) = (1 - \theta_i)C(i; n) + \theta_i\{C(i; n) + [1 - C(i; n)]\beta\Delta_i\}, \tag{6}$$

where

$$\Delta_i = \lambda\frac{1 - [(1-\lambda)(1-\beta)]^{r-i}}{1 - (1-\lambda)(1-\beta)} + [(1-\lambda)(1-\beta)]^{r-i}. \tag{7}$$

Inspection of (7) indicates that Δ, defined over list positions, is bounded between zero and unity. The function assumes a minimum at position one and increases as i becomes large to a maximum value of unity at position r.

The solution of difference equation (6) is

$$C(i; n) = 1 - [1 - C(i; 0)][1 - \theta_i\beta\Delta_i]^n \tag{8}$$

(cf. [5]).

Similar sets of equations (see Appendix A) can be written for the probability of an anticipatory error and failure-to-respond. However, for simplicity, analysis is limited here to $C(i; n)$.

For the typical rote serial learning situation, assume $C(i; 0) = 0$; that is, on the first run through the list S will make no correct anticipations. The probability of an error on trial n at position i is $[1 - C(i; n)]$, and the number

of errors at position i during the first $x + 1$ trials is

$$\sum_{n=0}^{x} [1 - C(i; n)] = \frac{1 - [1 - \theta_i\beta\Delta_i]^{x+1}}{\theta_i\beta\Delta_i}. \tag{9}$$

As x becomes large this expression approaches

$$1/(\theta_i\beta\Delta_i). \tag{10}$$

Application to Data

Data have been collected for different list lengths with a one-minute intertrial interval [1]. The lists were composed of familiar and easily pronounced two-syllable adjectives; no two words possessed similar meaning or phonetic construction. The data on total number of errors over the first 16 trials at each list position are presented in Figure 2. Each curve is based

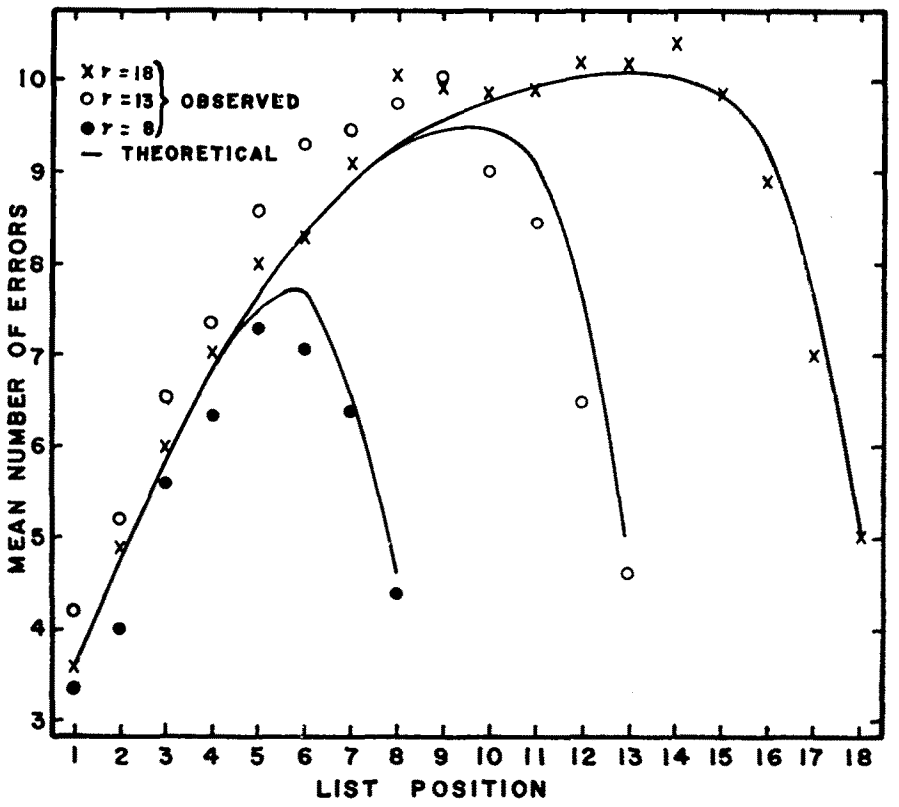

FIGURE 2

Theoretical and observed values of mean number of errors by serial positions over the first 16 trials for lists of length 8, 13, and 18.

on the records of 42 *S*s obtained in a situation employing a latin square design. Evidence on intertrial interval [1] suggests that the one-minute period experimentally approximates the theoretical infinite intertrial interval. Therefore equations (2) and (10) are applicable. These equations were employed to provide a visual fit to data for the list in which r equals 18; the obtained parameter values were $\lambda = .41$, $\beta = .55$, $\theta_1 = 1.00$, $c = .64$, and $\eta = .35$. These values were substituted in equations (2) and (10) to yield predicted curves for r equal to 8 and 13. An inspection of Figure 2 indicates close agreement between predicted and observed values.

Discussion

In the introduction the class of rote serial learning experiments to which the model is presumed to apply was delimited. The reasons for these restrictions are:

(*a*) *Moderate presentation rate.* A presentation rate that is too rapid would tend to decrease the likelihood of overt verbal responses and lead to an increase in the number of failures-to-respond. Consequently the model when applied to conditions of rapid presentation would underestimate the observed number of failures-to-respond. On the other hand, the model assumes that a single sample is drawn from S_i during the W_i exposure, an assumption which is to depend on a short exposure period. Experimentally these difficulties can be resolved by a short word exposure period followed by a blank exposure during which *S* provides an anticipation or failure-to-respond. An extension of the model to the case of a rapid rate has been examined, but the equations will not be displayed here.

(*b*) *Highly dissimilar words.* It is required in the model that the S_i sets be pairwise disjoint. This simplifying assumption is suspect for any serial learning situation, but it appears to provide an adequate approximation in this restricted situation. For the case of highly similar list words a set of elements common to each S_i would be introduced; the additional problems generated in this case are not considered here.

(*c*) *Familiar and easily pronounced words.* For the model, this restriction refers to a state such that the occurrence of the hypothesized W_i—R'_{i-1} relation is invariant over trials. For nonsense syllable learning the model would require, as an additional feature, a function describing the acquisition over trials of the W_i—R'_{i-1} connection [7].

In analyzing the model, the case where the intertrial interval is long has been considered. With a short interval the equations become more complex. Now some elements sampled on trial n remain available throughout the intertrial interval and into the next run through the list. For example, assume that an element is sampled from S_{r-1} on trial n and not returned to S^* for five h-intervals; the probability of this event is $\lambda(1 - \lambda)^4\theta_{r-1}$. When

$k = 1$, the element will be returned after the occurrence of R_i' on trial $n + 1$. Consequently, there is a probability $\beta[1 - C(r - 1; n)]$ that this element is conditioned to the response class R_1 . The element, when sampled again, increases the likelihood of an R_1 anticipatory response which, at position $r - 1$, would be classified as a perseverative error. It follows that the shorter the intertrial interval the greater the number of perseverative errors. This result has been experimentally verified [1].

Appendix A

Probability of a Failure-to-Respond and an Anticipatory Error

For the case of an infinite intertrial interval the probability of a failure-to-respond at position i on trial $n + 1$ is

$$\bar{R}(i; n + 1) = (1 - \theta_i)\bar{R}(i; n) + \theta_i[1 - C(i; n)](1 - \beta)\Delta_i \,. \tag{11}$$

The solution [5, p. 584] of this difference equation is

$$\bar{R}(i; n) = (1 - \theta_i)^n\bar{R}(i; 0) + \frac{(1 - \beta)\Delta_i}{1 - \beta\Delta_i}[(1 - \theta_i\beta\Delta_i)^n - (1 - \theta_i)^n], \tag{12}$$

where $\bar{R}(i; 0)$ is the probability of a failure-to-respond on the initial run through the list. The probability of an anticipatory error is

$$A(i; n) = 1 - C(i; n) - \bar{R}(i; n). \tag{13}$$

For the typical experimental situation, assume $C(i; 0) = 0$ and $\bar{R}(i; 0) = 1$; then (13) reduces to

$$A(i; n) = \frac{1 - \Delta_i}{1 - \beta\Delta_i}[(1 - \theta_i\beta\Delta_i)^n - (1 - \theta_i)^n]. \tag{14}$$

(12) and (14) when summed over the first x trials, as was done in (9) for incorrect responses, produce functions for failures-to-respond and anticipatory errors of the form reported by Deese and Kresse [2].

Appendix B

List of Symbols and Their Meanings

$A(i; n)$	probability of an anticipatory error at position i on trial n.
β	conditioning constant associated with an incorrect anticipation.
c	$\theta_1 - \theta_\infty$.
$C(i; n)$	probability of a correct anticipation at position i on trial n.
Δ_i	function defined over i; dependent on r, λ, and β.
η	decay constant related to the decrement in θ_i as i increases.
h	time of a single word exposure.
k	number of h-intervals in the intertrial interval.

λ	probability that an available element will be returned to S^* during the next h-interval.
n	number of trial.
r	list length.
R_i'	hypothesized covert response; reading W_{i+1}.
R_i	response class; overt anticipation of W_{i+1}.
$\bar{R}$	response class; failure-to-respond.
$R_i(i)$	R_i recorded by experimenter to W_i.
$\bar{R}(i; n)$	probability of a failure-to-respond at position i on trial n.
S^*	set of stimulus elements of which all S_i are subsets.
S_i	set of stimulus elements associated with W_i.
θ_i	probability of sampling an element from S_i when W_i occurs.
W_i	ith word presentation, where W_1 is cue for first anticipation.

REFERENCES

[1] Atkinson, R. C. An analysis of rote serial position effects in terms of a statistical model. Unpublished doctor's dissertation, Indiana Univ., 1954.
[2] Deese, J. and Kresse, F. H. An experimental analysis of the errors in rote serial learning. *J. exp. Psychol.*, 1952, **44**, 199-202.
[3] Estes, W. K. Toward a statistical theory of learning. *Psychol. Rev.*, 1950, **57**, 94-107.
[4] Estes, W. K. and Burke, C. J. A theory of stimulus variability. *Psychol. Rev.*, 1953, **60**, 276-286.
[5] Jordan, C. Calculus of finite differences. New York: Chelsea, 1950.
[6] McGeoch, J. A. and Irion, A. L. The psychology of human learning. New York: Longmans, Green, 1952.
[7] Noble, C. E. The effect of familiarization upon serial verbal learning. *J. exp. Psychol.*, 1955, **49**, 333-337.

Manuscript received 1/12/56
Revised manuscript received 2/22/56