PSYCHOMETRIKA—VOL. 26, NO. 3
SEPTEMBER, 1961

APPLICATION OF A MODEL TO PAIRED-ASSOCIATE LEARNING*

GORDON H. BOWER
STANFORD UNIVERSITY

The proposal is made to consider a paired-associate item as becoming conditioned to its correct response in all-or-none fashion, and that prior to this conditioning event the subject guesses responses at random to an unlearned item. These simple assumptions enable the derivation of an extensive number of predictions about paired-associate learning. The predictions compare very favorably with the results of an experiment discussed below.

This report describes an elementary model for the stimulus-response association process in paired-associate learning, displays an extensive number of derivations from the axioms of the model, and describes the agreement of the model with some experimental results. Paired-associate learning (PAL) as it is frequently studied involves two, at least conceptually, distinct processes: the learning of relevant responses to the general situation (e.g., as in nonsense syllable-syllable pairs), and the associative "hook-up" of these relevant responses to their appropriate stimulus members. In the belief that fractionating experimental problems leads to quicker understanding of the processes involved, this article is directed to only the second process listed above, the associative hook-up of relevant responses to their respective stimuli. The hope is that once this process is better understood the other problems, having to do with the learning of integrated response units in the situation, will become more amenable to experimental attack.

The way in which the response learning requirement was eliminated in the present experiments was to (i) use responses familiar to the subject, and (ii) inform him of the response alternatives before the experiment began. For these purposes, it was found that the first several integers $(1, 2, \cdots, N)$ worked admirably. Other responses meeting the above requirements could have been used, provided precautions were taken to prevent the subject from forgetting some of the available responses during the course of the experiment. The other procedural peculiarity of these experiments was the requirement that the subject make a relevant response to each stimulus item on each trial. This procedure necessarily involved permitting the subject to control his exposure time to each stimulus.

If there are K items in the list, then a "trial" will be defined as one cycle

*This research was supported by a grant, M-3849, from the National Institutes of Mental Health, United States Public Health Service.

of presentation of each of the K items, the order of appearance of the items being randomized over successive trials. Considering only a single stimulus item for a given subject, we may categorize his responses as correct or incorrect (or, 0 or 1, as we shall use later); over successive trials there will be some particular sequence of 1's and 0's to that item. Stripped to its barest essentials, the job for a theory of PAL is to describe and account for the general characteristics of these sequences. The best job of description, of course, would be to reproduce the original sequences. Theories, as economic abstractions, do not perform this task but they can provide general descriptions (e.g., the trial number of the second success) about a sample of sequences allegedly generated under the same process laws. Obviously, models that deliver predictions about many different aspects of such sequences are preferable to less tractable models, since each prediction provides an opportunity to test the adequacy of the model. In turn, the number of predictions derivable in closed form from a model reflects to a large extent the simplicity of the assumptions used to represent the process under consideration. The assumptions of the model to be presented appear to achieve almost maximal simplicity for a model about learning; accordingly, it is possible to derive in closed form an extensive number of predictions (theorems) referring to properties of the response sequences obtained from the learning subject.

The model to be described is derived within the general framework of a stimulus sampling theory of learning [9] but with the assumption that each experimental source of stimulation (i.e., the stimulus member of a paired-associate item) may be represented by a small number of stimulus components or elements. The original investigation of small-element learning models began with a paper by Estes [10] and has been carried on by a number of people. Suppes and Atkinson [15] give an extensive development of such models and show their application to a variety of learning experiments. In the initial development of stimulus sampling theory [8, 9] it was assumed that the population of stimulus components from which the subject sampled on each trial was large. Since conditioning was assumed to take place only with respect to the sampled elements, the model implied relatively gradual changes over trials in the proportion of conditioned elements in the population and hence in response probability. Recent developments with small-element models differ in that the population of stimulus elements is assumed to be small (e.g., one or two elements) so that response probability may take on only a few values over the course of a learning experiment. The common assumption is that only one of these stimulus elements may be sampled on each trial and that the sampled element becomes conditioned to the reinforced response with probability c on every trial. Besides considerable simplification of the mathematics of stimulus sampling theory, the small-element assumptions deliver some predictions which differ markedly from the large-element (i.e., linear) model assumptions; some of these differences are noted and will be compared with data.

The basic notion of the present model is that each stimulus item in the list of paired associates may be represented by exactly one stimulus element within the model and that the correct response to that item becomes associated in all-or-none fashion. Considering only a single item, it can be in either of two "states" on each trial: conditioned or not conditioned to the correct response. The effect of a reinforced trial (i.e., evoking the correct response in the presence of the stimulus item) is to provide an opportunity for the item to become conditioned. The single parameter of the model is c, the probability that an unconditioned item will become conditioned as the result of a reinforced trial. All items begin in the unconditioned state; the effect of continued reinforced trials is to provide repeated opportunities for the item to become conditioned.

If the item has become conditioned, then continued reinforcements of the same correct response will ensure that the item remains conditioned. The probability of the correct response when the item is conditioned is unity. The probability of the correct response when the item is not conditioned depends upon the exact experimental procedure used. In experiments by the writer, the subjects were told the N responses (integers 1, 2, $\cdots$, N) available to them and were told to respond on every trial regardless of whether they knew the correct number. If the N numbers occur equally often as the to-be-learned responses to the items, then the probability that the subject will guess correctly on an unlearned item is $1/N$; correspondingly, his probability of guessing incorrectly is $1 - (1/N)$. Our discussion of the one-element model is oriented specifically towards such an experimental procedure.

Because of the way the model is formulated, there is a partial determinism between the response sequence and the sequence of conditioning states. Specifically, if the subject responds incorrectly to a given item on trial n, then that item was not in the "conditioned" state on trial n. This feature is very helpful in deriving a number of the theorems about errors. If the subject responds correctly, however, then we cannot uniquely specify his state of conditioning, since he may have guessed correctly. Thus, it is not a consequence of the model that the subject's first correct response will be followed with probability one by correct responses on subsequent trials.

After working with the latter model for some time, it came to the writer's attention that Bush and Mosteller [6] had previously published a model for "one-trial learning" that is almost identical to the one stated above. Thus, there can be no pretense to priority in the current formulation of these elementary notions about the learning process. The present account does go beyond the abbreviated discussion by Bush and Mosteller in deriving a large number of predictions from the model and in applying the theory with some success to verbal learning. Although their approach and the present one differ slightly in assumptions about initial conditions, the derivational techniques are sufficiently similar so that theorems can be transposed, with appropriate modifications, from one system to the other. [According to the Bush and Mosteller assumptions, a proportion c of the response sequences (subjects or items) begin in the conditioned state, and this same value of c is assumed to be the learning rate constant.]

Throughout the following sections, the predictions derived from the model will be compared with data from an experiment which now will be described. Twenty-nine subjects learned a list of ten items to a criterion of two consecutive errorless cycles. The stimuli were different pairs of consonant letters; the responses were the integers 1 and 2, each response assigned as correct to a randomly selected five stimuli for each subject. A response was obtained from the subject on each presentation of an item and he was informed of the correct answer following his response. The deck of ten stimulus cards was shuffled between trials to randomize the presentation order of the stimuli.

Axioms and Theorems about Total Errors

Axioms

1. *Each item may be represented by a single stimulus element which is sampled on every trial.*
2. *This element is in either of two conditioning states:* C_1 *(conditioned to the correct response) or* C_0 *(not conditioned).*
3. *On each reinforced trial, the probability of a transition from* C_0 *to* C_1 *is a constant,* c*; the probability of a transition from* C_1 *to* C_1 *is one.*
4. *If the element is in state* C_1 *then the probability of a correct response is one; if the element is in state* C_0 *, then the probability of a correct response is* $1/N$*, where* N *is the number of response alternatives.*
5. *The probability* c *is independent of the trial number and the outcomes of preceding trials.*

The trial to trial sequence of conditioning states forms a Markov chain, with C_1 being an absorbing state. The transition probabilities are given in the following matrix.

$$P = \begin{array}{c|cc} & C_1 & C_0 \\ \hline C_1 & 1 & 0 \\ C_0 & c & 1 - c. \end{array} \tag{1}$$

It is easy to show that the nth power of the transition matrix is

$$(2) \qquad P^n = \begin{array}{c|cc} & C_1 & C_0 \\ \hline C_1 & 1 & 0 \\ C_0 & 1 - (1 - c)^n & (1 - c)^n. \end{array}$$

We explicitly assume that all items start out in state C_0 (i.e., are not conditioned initially). Thus, starting out in state C_0, the probability of still being in state C_0 after n reinforced trials is $(1 - c)^n$, which approaches zero as n becomes large. Thus, for $c > 0$, with probability one the process will eventually end in conditioning state C_1 (i.e., will become conditioned).

For each item, define a sequence of response random variables, x_n, which take on the value 1 if an error occurs on trial n, or the value 0 if a success occurs on n. From the axioms, the conditional probabilities of an error given states C_1 or C_0 at the beginning of trial n are

$$(3) \qquad \Pr\{x_n = 1 \mid C_{1,n}\} = 0 \quad \text{and} \quad \Pr\{x_n = 1 \mid C_{0,n}\} = 1 - \frac{1}{N}\cdot$$

To obtain the average probability of an error on the nth trial, q_n, multiply these conditional probabilities by the probabilities of being in C_1 or C_0, respectively, at the start of trial n:

$$\begin{aligned}(4) \qquad q_n = \Pr\{x_n = 1\} &= \Pr\{x_n = 1 \mid C_{1,n}\} \Pr\{C_{1,n}\} \\ &\quad + \Pr\{x_n = 1 \mid C_{0,n}\} \Pr\{C_{0,n}\} \\ &= 0 + \left(1 - \frac{1}{N}\right)(1 - c)^{n-1} = \left(1 - \frac{1}{N}\right)(1 - c)^{n-1}.\end{aligned}$$

The expected total number of errors, u_1, before perfect learning is given by

$$(5) \quad u_1 = E\left[\sum_{n=1}^{\infty} x_n\right] = \sum_{n=1}^{\infty} \Pr\{x_n = 1\} = \sum_{1}^{\infty} \left(1 - \frac{1}{N}\right)(1 - c)^{n-1} = \frac{1 - \frac{1}{N}}{c}.$$

The expected total errors per item serves as a stable estimator of c. For the experiment described above with $N = 2$, the average number of errors per item was 1.45. Equating u_1 in (5) to 1.45, the c value obtained is .344. This estimate of c will be fixed throughout the remaining discussion of these data. Using this value of c in (4), the predicted learning curve in Fig. 1 is obtained.

In the expression for u_1, all errors are weighted equally. It is also possible to derive expressions for various weighted sums of errors, as Bush and Sternberg [7] have shown for the linear model. The results here are identical with

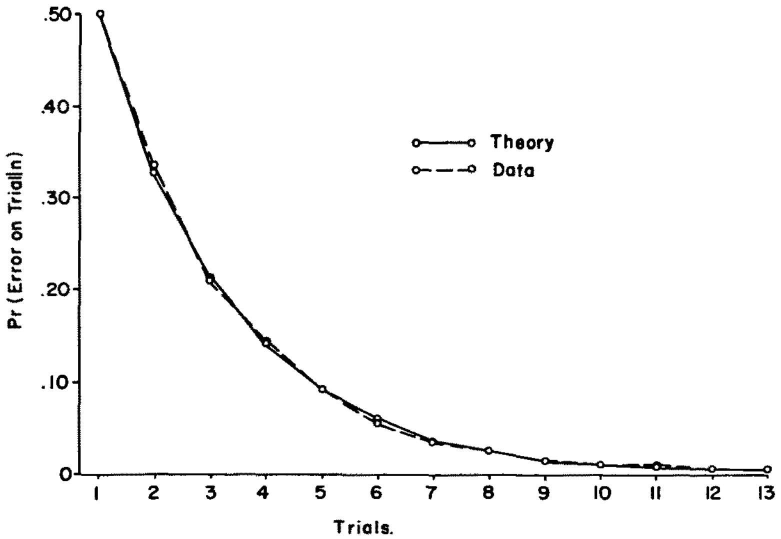

FIGURE 1

q_n, the Probability of an Incorrect Response over Successive Trials of the Experiment

their results. Three examples of the expectation of weighted error sums are given below.

$$E\left[\sum_{n=1}^{\infty} nx_n\right] = \sum_{n=1}^{\infty} n\left(1 - \frac{1}{N}\right)(1 - c)^{n-1} = \frac{1 - \dfrac{1}{N}}{c^2} = \frac{u_1}{c}\ ; \tag{6}$$

$$E\left[\sum_{n=1}^{\infty} \frac{x_n}{n}\right] = \sum_{n=1}^{\infty} \frac{\left(1 - \dfrac{1}{N}\right)(1 - c)^{n-1}}{n} = \frac{1 - \dfrac{1}{N}}{1 - c} \log \frac{1}{c}\ ; \tag{7}$$

$$E\left[\sum_{n=1}^{\infty} \frac{x_n}{(n - 1)!}\right] = \left(1 - \frac{1}{N}\right) \sum_{m=0}^{\infty} \frac{(1 - c)^m}{m!} = \left(1 - \frac{1}{N}\right)e^{1-c}. \tag{8}$$

It is possible to obtain the distribution of the total number of errors on each item. This distribution was derived by Bush and Mosteller; their result is readily translated into the terms of the current approach to the theory. If we let T represent the total number of errors (to perfect learning) on a single item, the probability distribution of T is

$$\Pr\{T = k\} = \begin{cases} b/N & \text{for} \quad k = 0 \\ \dfrac{b(1 - b)^k}{1 - c} & \text{for} \quad k \geq 1, \end{cases} \tag{9}$$

where

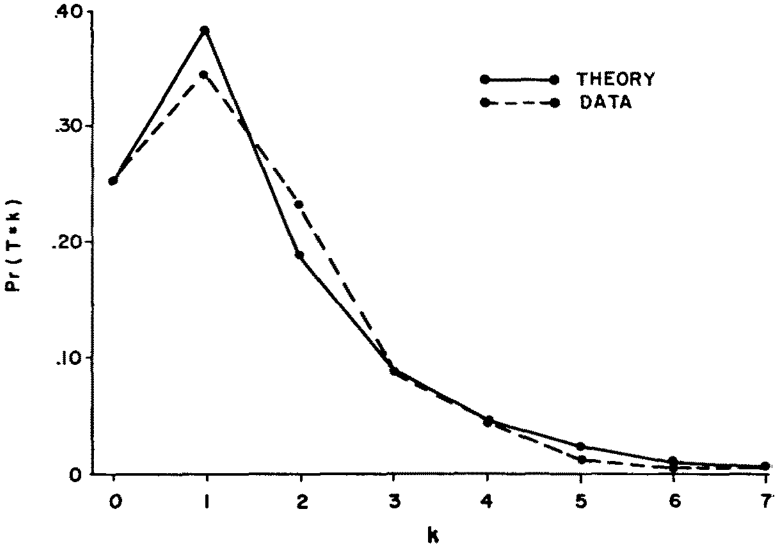

FIGURE 2
Distribution of T, the Total Number of Errors per Item

$$b = \frac{c}{1 - \frac{1-c}{N}}.$$

The mean of T was derived as u_1 in (5); the variance of T is given by

$$\text{Var}\,(T) = u_1 + (1 - 2c)u_1^2\,. \tag{10}$$

The predicted and obtained distributions of T are shown in Fig. 2.

Sequential Properties of the Model

Predictions about sequential features of the data may be obtained by considering runs of errors. To date only mean values of the various run distributions have been derived; higher moments will not be discussed. Let r_j represent the number of error runs of length j in an infinite number of trials; we seek the expectation of r_j. For these purposes, it is convenient to define another random variable, u_j, which counts the number of j-tuples of errors that occur in an infinite sequence of trials. Formally, define u_j as

$$u_j = \sum_{n=1}^{\infty} x_n x_{n+1} \cdots x_{n+j-1} \quad \text{for} \quad j = 1, 2, \cdots. \tag{11}$$

The product, $x_n x_{n+1} \cdots x_{n+j-1}$, has the value one only when j consecutive errors occur starting with the error on trial n. It may be seen that u_1 is just the total number of errors.

To make clear how the u_j are being counted and their relation to the r_j, consider the possible sequence

$$111110011000110100 0 \cdots \text{ (all the rest zeros).}$$

For this sequence,

$$u_1 = 10, \quad u_2 = 6, \quad u_3 = 3, \quad u_4 = 2, \quad u_5 = 1;$$

$$r_1 = 1, \quad r_2 = 2, \quad r_3 = r_4 = 0, \quad r_5 = 1, \quad \underline{R = \sum_j} r_j = 4.$$

R is the total number of error runs. In an excellent article, Bush [5] has shown that the r_j can be expressed as linear combinations of the u_j. In particular,

$$(12) \qquad r_j = u_j - 2u_{j+1} + u_{j+2},$$

and

$$(13) \qquad R = \sum_{j=1}^{\infty} r_j = u_1 - u_2 .$$

Having expressed the r_j in terms of the u_j, we now turn to deriving from the model the expected value of u_j. We proceed as follows:

$$(14) \qquad E(u_j) = E\left[\sum_{n=1}^{\infty} x_n \cdot x_{n+1} \cdots \cdot x_{n+j-1}\right] = \sum_{n=1}^{\infty} \Pr\{x_n = 1\} \cdot \Pr\{x_{n+1} = 1 \mid x_n = 1\} \Pr\{x_{n+2} = 1 \mid x_n \cdot x_{n+1} = 1\} \cdots \cdot \Pr\{x_{n+j-1} = 1 \mid x_n \cdot x_{n+1} \cdots x_{n+j-2} = 1\}.$$

Because of the Markovian properties of the model, the lengthy conditional probabilities on the right-hand side can be simplified, viz.,

$$(15) \qquad \Pr\{x_{n+i} = 1 \mid x_n = 1, x_{n+1} = 1, \cdots, x_{n+i-1} = 1\} = \Pr\{x_{n+i} = 1 \mid x_{n+i-1} = 1\}.$$

That is, if the subject made an error on the preceding trial, then that is all the information there is to be extracted from the entire preceding sequence of responses. His error tells us that his conditioning state on the preceding trial was C_0; the probability of an error on the current trial is then

$$(16) \qquad \Pr\{x_{n+1} = 1 \mid x_n = 1\} = c \cdot 0 + (1 - c)\left(1 - \frac{1}{N}\right) = (1 - c)\left(1 - \frac{1}{N}\right) = \alpha,$$

and, moreover, this holds for any trial number n. Thus, using relations (15) and (16), the equation for u_j becomes

$$(17)\qquad E(u_j) = \sum_{n=1}^{\infty} \Pr\{x_n = 1\} \Pr\{x_{n+1} = 1 \mid x_n = 1\} \cdots \cdot \Pr\{x_{n+j-1} = 1 \mid x_{n+j-2} = 1\}$$

$$= \sum_{n=1}^{\infty} \Pr\{x_n = 1\} \underbrace{\alpha \cdot \cdots \cdot \alpha}_{(j-1)\text{ times}}.$$

$$E(u_j) = \alpha^{j-1} \sum_{n=1}^{\infty} \Pr\{x_n = 1\} = u_1 \alpha^{j-1}.$$

With these values in hand, now calculate R and r_j, using relations (12) and (13).

$$(18)\qquad E(R) = E(u_1) - E(u_2) = u_1(1 - \alpha),$$

$$(19)\qquad E(r_j) = E(u_j) - 2E(u_{j+1}) + E(u_{j+2}) = u_1(1 - \alpha)^2 \alpha^{j-1}$$

$$= R(1 - \alpha)\alpha^{j-1}.$$

Another useful summary of sequential properties in the data is the extent to which an error on trial n tends to be followed by an error k trials later, without regard to what responses intervene between trials n and $n + k$. Define $c_{k,n}$ as $x_n \cdot x_{n+k}$; this expression will have the value 1 only if errors occur on both trials n and $n + k$. It may be noted that $c_{k,n}$ summarizes the same features as does an autocorrelation of lag k. The expectation of $c_{k,n}$ is

$$(20)\qquad E(c_{k,n}) = E(x_n \cdot x_{n+k}) = E(x_{n+k} \mid x_n) \cdot E(x_n)$$

$$= \Pr\{x_{n+k} = 1 \mid x_n = 1_j \Pr\{x_n = 1\}.$$

To find the conditional probability above, note that for an error to occur on trial $n + k$ it must be the case that conditioning has failed to occur during the intervening k trials, and moreover that the subject guesses incorrectly on trial $n + k$. The probability of this joint event is

$$(21)\qquad \Pr\{x_{n+k} = 1 \mid x_n = 1\} = (1 - c)^k \left(1 - \frac{1}{N}\right).$$

Therefore,

$$(22)\qquad E(c_{k,n}) = \left(1 - \frac{1}{N}\right)(1 - c)^k \left(1 - \frac{1}{N}\right)(1 - c)^{n-1}.$$

A convenient statistic for comparison with data is obtained by taking the "autocorrelation" of x_n and x_{n+k} over all trials n of the experiment. Defining c_k as the mean value of this random variable,

$$(23)\qquad c_k = E\left[\sum_{n=1}^{\infty} x_n x_{n+k}\right] = \sum_{n=1}^{\infty} E(c_{k,n}) = u_1\left(1 - \frac{1}{N}\right)(1 - c)^k$$

$$\text{for}\quad k = 1, 2, 3, \cdots .$$

Predicted and observed values of c_1, c_2, and c_3 are given in Table 1.

It is a simple matter to construct other statistics which capture various features of the sequential dependencies in the response sequence. Such statistics are expressible as various sums and/or products of the x_n. One illustration will be provided here to demonstrate the general derivational techniques. In order to predict the average number of alternations of successes and failures that occur over the response sequence, define a random variable A_n which will count an alternation between trials n and $n + 1$. Hence,

$$(24)\qquad A_n = (1 - x_n)x_{n+1} + x_n(1 - x_{n+1}).$$

It will be noted that A_n takes on the value 1 either if a success occurs on trial n and a failure on trial $n + 1$ or if a failure occurs on n and a success on $n + 1$. Multiplying out and taking the expectation of A_n yields

$$(25)\qquad E(A_n) = \frac{\alpha}{N}(1 - c)^{n-1} + (1 - \alpha)\left(1 - \frac{1}{N}\right)(1 - c)^{n-1}.$$

The average of the sum of A_n over trials is

$$(26)\qquad A = E[\sum A_n] = \underline{u_1\Big[c + \frac{2(1 - c)}{N}}\Big] .$$

Errors during Various Parts of Learning

In this section we derive the distribution of the number of errors between the kth and $(k + 1)$st success and also of the number of errors between the kth and $(k + 2)$nd success. As special cases of these general results, for $k = 0$ we obtain the distributions of errors before the first and before the second success. The methods employed in these derivations are general so that the distribution of errors between the kth and $(k + m)$th success could be obtained, the sole limitation being that the expressions get progressively more cumbersome as m is increased.

Consider first the distribution of the number of errors occurring between the kth and $(k + 1)$st success. Let J_k be this random variable; it can take on the values 0, 1, 2, $\cdots$ of the non-negative integers. Errors following the kth success can occur only if the kth success itself came about by guessing (rather than via prior conditioning). Thus, the probability that the kth success occurred by guessing (call it g_k) will play a central role in the expression for the distribution of J_k. To forego for the moment the derivation of g_k, write the distribution of J_k as

$$\Pr\{J_k = i\} = \begin{cases} 1 - g_k\alpha & \text{for} \quad i = 0 \\ g_k(1 - \alpha)\alpha^i & \text{for} \quad i > 0. \end{cases} \tag{27}$$

For example, the probability of three errors between the kth success and the next one is given by the joint probability of (i) the kth success occurred by guessing, (ii) conditioning failed to occur at the end of trials, k, $k + 1$, and $k + 2$ and incorrect guesses occurred on trials $k + 1$, $k + 2$, and $k + 3$, the probability of this joint event being $(1 - c)^3\ (1 - 1/N)^3 = \alpha^3$, and (iii) given that the element was not conditioned at the start of trial $k + 3$, a correct response occurs on trial $k + 4$ with probability $1 - \alpha$. To obtain the term for $J_k = 0$, note that no errors could occur either if the kth success occurred via prior conditioning (with probability $1 - g_k$) or, having guessed the kth success, a success occurs on the next trial with probability $1 - \alpha$. The sum of these two terms, $1 - g_k$ and $g_k(1 - \alpha)$, gives the probability that $J_k = 0$.

From the distribution in (27) one obtains the mean and variance of J_k.

$$E(J_k) = \frac{a g_k}{1 - \alpha}, \qquad \mathrm{Var}\,(J_k) = \frac{\alpha g_k}{(1 - \alpha)^2}\,[1 + \alpha(1 - g_k)]. \tag{28}$$

The task now is to derive g_k, the probability that the kth success occurs by guessing. Consider g_1, the probability that the first success occurs by guessing. It is

$$\begin{aligned} g_1 &= \frac{1}{N} + \left(1 - \frac{1}{N}\right)(1 - c)\,\frac{1}{N} + \left(1 - \frac{1}{N}\right)^2 (1 - c)^2\,\frac{1}{N} + \cdots \\ &= \frac{1}{N} \sum_{j=0}^{\infty} \alpha^i = \frac{1}{N(1 - \alpha)}. \end{aligned} \tag{29}$$

That is, the subject guesses correctly on the first trial with probability $1/N$; he may fail there so the item does not become conditioned and he guesses correctly on the second trial, and so on. It can be shown for $k > 1$ that a general recursion holds for g_k, viz.,

$$g_k = g_{k-1}(1 - c)\left[\frac{1}{N} + \alpha\,\frac{1}{N} + \alpha^2\,\frac{1}{N} + \cdots\right] = g_{k-1}(1 - c)g_1\,. \tag{30}$$

That is, for the kth success to occur by guessing, it must be the case that (i) the $(k - 1)$st success occurred by guessing, (ii) conditioning failed to occur on the trial of the $(k - 1)$st success, and (iii) starting out not conditioned on the next trial, the next correct response also occurs by guessing, with probability g_1.

Equation (30) is a standard linear difference equation having the solution

$$g_k = (1 - c)^{k-1} g_1^k = g_1\left(1 - \frac{c}{1 - \alpha}\right)^{k-1}. \tag{31}$$

Since $c < 1 - \alpha$, it follows that g_k decreases exponentially with k. This result is intuitively clear: the tenth success is less likely to occur by guessing than is, say, the second success. Corresponding to the decrease in g_k, the average errors between the kth and $(k + 1)$st success is decreasing exponentially over k, as (28) shows.

We have been considering J_k for $k > 0$. The interpretation of J_0 is the number of errors before the first success. It is convenient to define g_0 as $1/(1 - c)$, although g_0 itself has no physical interpretation. Defining g_0 in this way, then the distribution of J_0, the errors before the first success, is given by (27). The distribution of J_0 has more intuitive appeal when written as

$$\Pr\{J_0 = i\} = \begin{cases} \dfrac{1}{N} & \text{for } i = 0 \\ \left(1 - \dfrac{1}{N}\right)(1 - \alpha)\alpha^{i-1} & \text{for } i \geq 1, \end{cases} \tag{32}$$

although formally it is the same as (27) with $g_0 = 1/(1 - c)$.

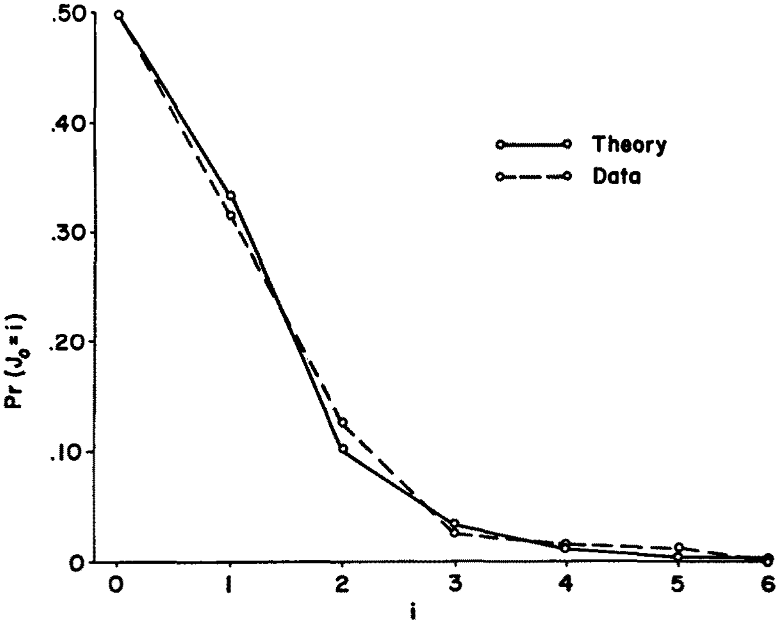

FIGURE 3
Distribution of J_0, the Number of Errors Before the First Success

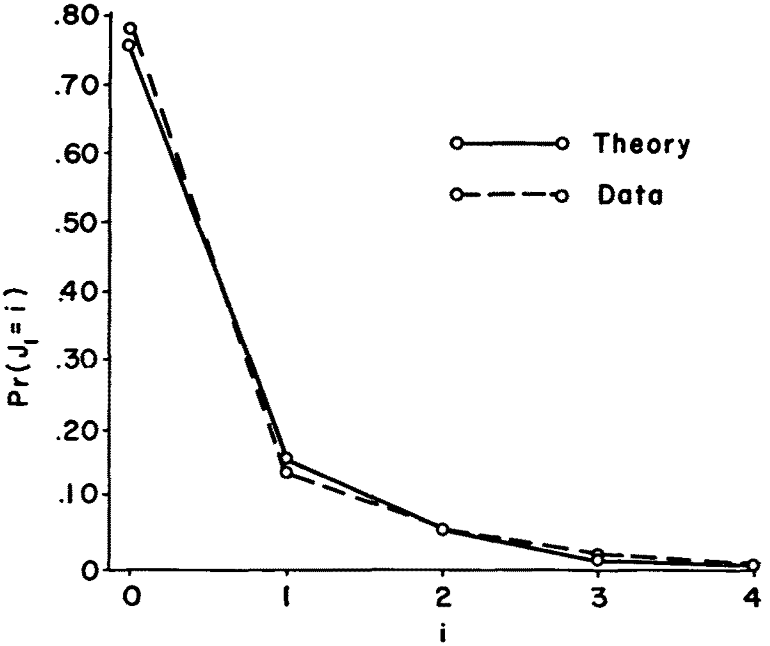

FIGURE 4
Distribution of J_1, the Number of Errors Between the First and Second Success

To illustrate the fit of the model to data, the distribution of the number of errors before the first success is shown in Fig. 3, and the mean and standard error, predicted and observed, are shown in Table 1. Also, the theoretical and observed distributions of J_1 , the number of errors between the first and second success, are shown in Fig. 4.

Using the J_k values so calculated, one obtains an expression for the average errors before the kth success. If F_k is defined as the cumulative errors before the kth success, then the obvious recursion on the means is

$$E(F_{k+1}) = E(F_k) + E(J_k). \tag{33}$$

The solution of this difference equation is

$$E(F_k) = \sum_{i=0}^{k-1} E(J_i) = \frac{\alpha}{1-\alpha} \sum_{i=0}^{k-1} g_i \,. \tag{34}$$

Substituting the values for g_i , the summation yields

$$\begin{aligned} E(F_k) &= \frac{1 - \frac{1}{N}}{1-\alpha} + \frac{\alpha g_1}{1-\alpha} \frac{[1 - (g_1(1-c))^{k-1}]}{[1 - g_1(1-c)]} \\ &= u_1 - \frac{\alpha}{Nc(1-\alpha)} [g_1(1-c)]^{k-1}, \end{aligned} \tag{35}$$

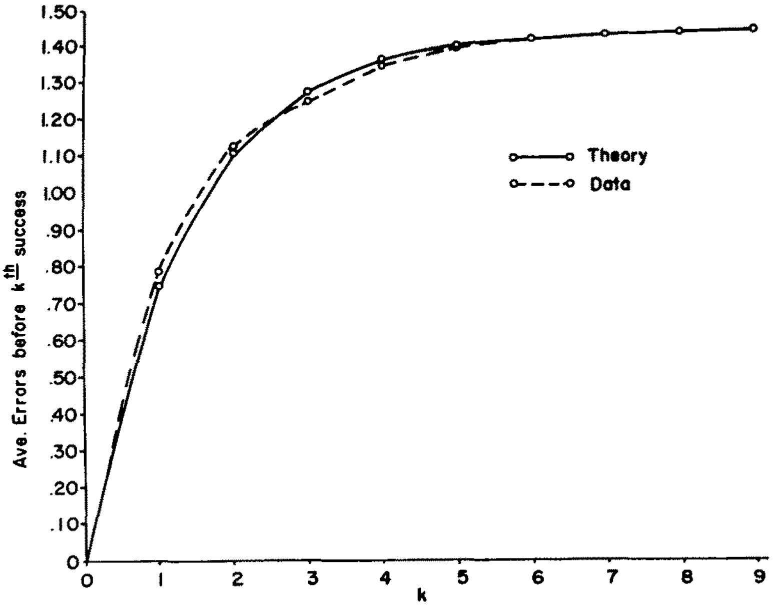

FIGURE 5

$E(F_k)$, the Expected Number of Errors Before the kth Success

where u_1 is the average total errors per item as given in (5). Equation (35) establishes the expected result that, for large k, the average number of errors before the kth success approaches the average total number of errors per item. In Fig. 5, the observed and predicted values of $E(F_k)$ through the ninth success are shown.

The distribution of the number of errors between the kth and $(k + 2)$nd success has been obtained and is presented here for completeness. Define S_k as the number of errors between the kth and $(k + 2)$nd success; it is clear that $S_k = J_k + J_{k+1}$. By specialization for $k = 0$, S_0 gives the distribution of the number of errors before the second success. The distribution of S_k , which is given here without proof (see [3]), is

$$(36)\qquad \Pr\{S_k = i\} = \begin{cases} 1 - g_k + g_k\left[c + \dfrac{(1-c)(1-\alpha)}{N}\right] & \text{for } i = 0 \\ g_k\left[c + \dfrac{(1-c)(1-\alpha)}{N}(i+1)\right]\alpha^i & \text{for } i \geq 1, \end{cases}$$

and the first and second raw moments of the distribution are

$$(37)\qquad E(S_k) = \frac{\alpha g_k}{(1-\alpha)^2}\left[c + \frac{2(1-c)}{N}\right],$$

$$E(S_k^2) = \frac{\alpha g_k}{(1-\alpha)^3}\,[2(1-\alpha)(1+2\alpha) - c(1+3\alpha)].$$

The g_k are as given before. Again, defining $g_0 = 1/(1 - c)$, (36) gives the distribution of the number of errors before the second success. The observed and predicted distributions of S_0 are shown in Fig. 6, and the mean and standard error, predicted and observed, are given in Table 1.

The preceding derivations have been carried out for the number of errors before the kth success, etc. The number of *trials* before the kth success is obviously related by a constant. Thus, the trial number of the kth success is the number of errors before the kth success, F_k , plus k. Changing to a "trial" notation shifts the origin (adds a constant) but does not affect the form or variance of the distribution.

The Trial Number of the Last Failure

Our purpose in this section is to derive the distribution of the trial number of the last error in an effectively infinite sequence of trials. However, before proceeding with this derivation, it is helpful to consider another statistic: the proportion of items characterized by having no errors following the first success. In the experimental data, a considerable percentage (62.8 percent, in fact) of the item protocols displayed this characteristic and the question arose whether the model would predict such results. Let p_1 represent the probability that a response sequence will exhibit this property of no errors following the first success. If b represents the probability of no more errors following a correct guess, then an expression for p_1 is

$$p_1 = 1 - g_1 + g_1 b = 1 - g_1(1 - b). \tag{38}$$

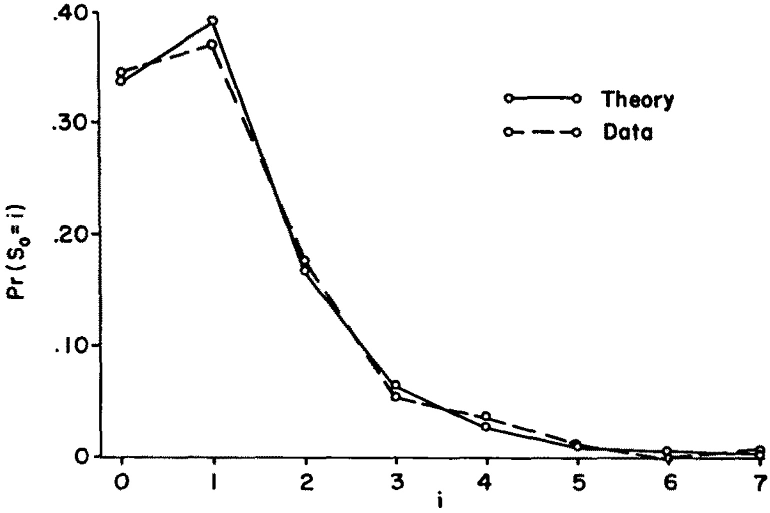

FIGURE 6
Distribution of S_0, the Number of Errors Before the Second Success

That is, a proportion $1 - g_1$ of the first correct responses come about via *prior conditioning* (so no more errors will occur), while $g_1 b$ represents the probability that the first correct response occurs by guessing but no more errors occur. To complete this derivation, b, the probability that no errors occur following a correct guess is

$$b = c + (1 - c)\frac{1}{N}c + (1 - c)^2\left(\frac{1}{N}\right)^2 c + \cdots \tag{39}$$

$$= \frac{c}{1 - \dfrac{(1 - c)}{N}} = \frac{c}{\alpha + c}.$$

That is, with probability c the item was conditioned on the trial on which the correct guess occurred; with probability $1 - c$ conditioning failed to occur on that trial, the subject guessed correctly on the next trial with probability $1/N$ and the item became conditioned then with probability c, and so on. This value of b is the same as that used in the distribution of the number of errors given in (9).

Substituting this result for b into (38),

$$p_1 = 1 - \frac{g_1\alpha}{\alpha + c}. \tag{40}$$

Using the estimate of c obtained earlier, the predicted p_1 is .638, which is quite close to the observed proportion of .628.

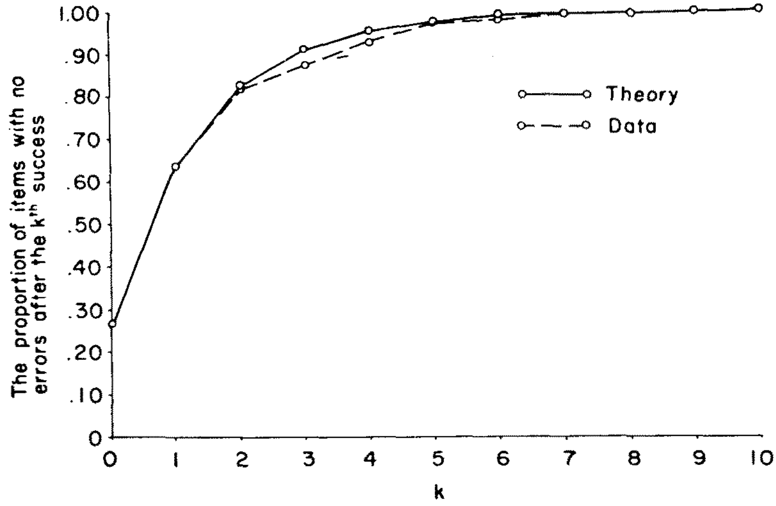

FIGURE 7

p_k, the Probability of Zero Errors Following the kth Success

As (40) suggests, define p_k to be the probability that there are no errors following the kth success. Using our previous result for g_k, one derives

$$(41) \qquad p_k = 1 - \frac{\alpha g_k}{\alpha + c} = 1 - \frac{\alpha g_1}{\alpha + c} [g_1(1 - c)]^{k-1}.$$

Observed and predicted values of p_k are shown in Fig. 7.

To determine the position of the last error, define n' as the random variable representing the trial number on which the last error occurs in an infinite sequence of trials. If no errors occur at all, then n' is set equal to zero. The probability distribution of n' is

$$(42) \qquad \Pr\{n' = k\} = \begin{cases} \dfrac{b}{N} & \text{for } k = 0 \\ b\left(1 - \dfrac{1}{N}\right)(1 - c)^{k-1} & \text{for } k \geq 1. \end{cases}$$

The first value is just $\Pr\{T = 0\}$, which was given in (9). If some errors occur, then for the last error to occur on trial k it must be the case that conditioning failed to occur on the preceding $k - 1$ trials, an incorrect guess occurred on trial k, but no errors followed that, with probability b. The mean and variance of n' are

$$(43) \qquad E(n') = m = \frac{b\left(1 - \frac{1}{N}\right)}{c^2} = \frac{bu_1}{c},$$

$$\operatorname{Var}(n') = m\left[\frac{2}{c} - 1 - m\right].$$

The observed and predicted distributions of n' are shown in Fig. 8; the mean and standard error, observed and predicted, are given in Table 1.

Consider now the distribution of the number of successes that intervene between the kth and $(k + 1)$st error, provided that a $(k + 1)$st error occurs. Because an error effectively "resets" the process to state C_0, the distribution of this random variable will be independent of k and of the trial number on which the leading error occurs. Let H represent the number of intervening successes. The distribution of H is given by

$$(44) \quad \Pr\{H = j\} = \frac{\alpha}{1 - b}\left[\frac{1 - c}{N}\right]^j = (\alpha + c)(1 - \alpha - c)^j, \; j = 0, 1, 2, \cdots.$$

The division by $1 - b$ establishes the condition that at least one more error will occur. The probability that the next error occurs on the very next trial is just $(1 - c)(1 - 1/N) = \alpha$; the probability that the next response is a correct guess and the error occurs on the following trial is $(1/N)(1 - c)\alpha$, and so on. Although the derivation of the number of successes before the

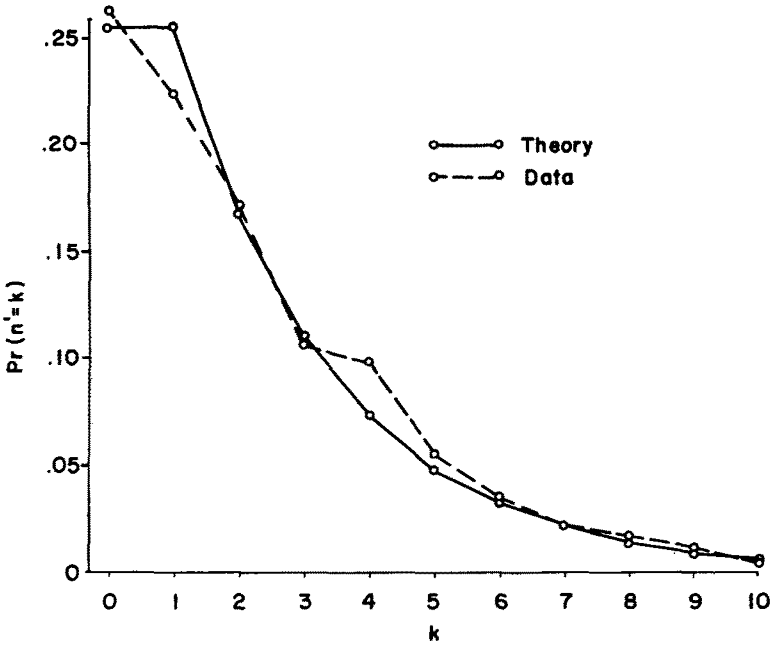

FIGURE 8
Distribution of n', the Trial Number of the Last Failure

first error (provided there is one) proceeds somewhat differently, the resulting distribution is identical to the distribution given in (44). The mean and variance of H are

$$E(H) = \frac{1 - \alpha - c}{\alpha + c}, \qquad \text{Var}(H) = \frac{1 - \alpha - c}{(\alpha + c)^2}. \tag{45}$$

The observed and predicted distributions of H are shown in Fig. 9, and the means and standard errors are given in Table 1.

The preceding analyses have been carried out for the responses to a single item over trials. If the items can be considered homogeneous in difficulty so that each learning process may be characterized by the same c value, then it is possible to derive a number of predictions about performance across items within a particular trial. If there are K items, then a run through the list (a trial) provides a sequentially ordered sample of size K from a binomial population. With this characterization, one can then derive various quantities of experimental interest, e.g., the probability that a perfect recitation of the list occurs on trial n, the average number of error runs of length j considering the K sequential samples on each trial, and so on. Derivations of such results are relatively easy and are presented in [3].

Comparison with Linear Model

What has been accomplished in the preceding sections is a detailed analysis of the sequence of response random variables. In terms of sheer bulk of predictions derivable from learning axioms, the sole comparable alternative is the single-operator linear model explored extensively by Bush and Sternberg [7]. It would be instructive, therefore, to place on record a detailed quantitative comparison of the fit of these two models to the present data. The basic notion of the linear model is that the associative strength between a stimulus and its correct response increases by a linear transformation following each reinforced trial. Stated differently, the probability of an error is expected to decrease by a constant fraction following each reinforced trial. If the initial error probability is $1 - (1/N)$ then over successive reinforced trials the error probability decreases, taking on a number of values intermediate between $1 - (1/N)$ and 0. In contrast, the one-element model proposed here assumes that the error probability has only two values, $1 - (1/N)$ or 0, and jumps from the first to the second value following the trial on which the all-or-none association is formed.

Although these conceptions differ markedly, the two models predict the same average learning curve. Thus, finer details of the data are required to differentiate these models. Since, according to the linear model, q_n decreases by the same fraction every trial, the response random variables, x_n ,

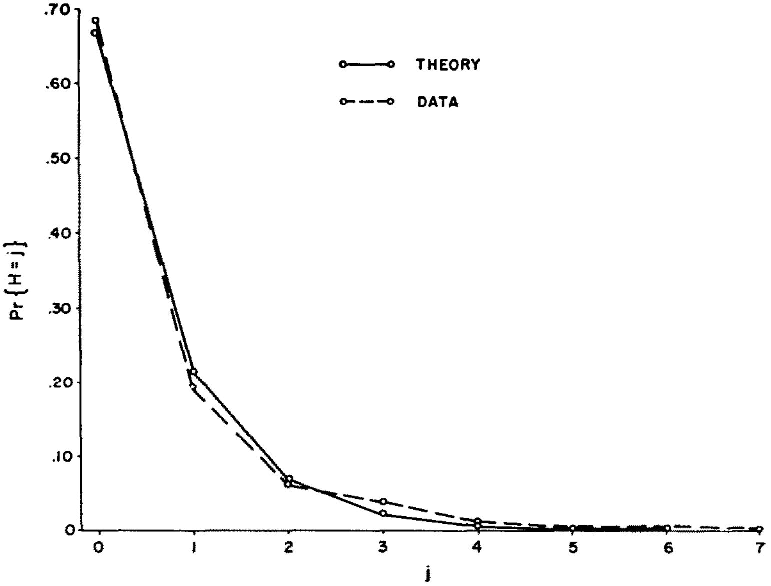

FIGURE 9
Distribution of H, the Number of Successes Intervening Between Adjacent Errors

are statistically independent; that is, the probability of an error on trial n is expected to be the same whether the subject responded correctly or incorrectly on the preceding trial. For the one-element model the x_n are not independent; whether we expect an error on trial n depends heavily on whether or not an error occurred on the preceding trial. Noting these differences, we are led to expect that the two models would be differentiated best by their predictions about sequential aspects of the data. Indeed this is the case, as may be seen in Table 1 which collects 19 comparisons of the one-element and linear models with data. The linear model predictions were obtained by referring to the theorems derived by Bush and Sternberg [7]. Three other

TABLE 1

Comparison of One-Element and Linear Models with Data

	Statistic	One element	Data	Linear
1.	Ave. errors per item	---	1.45	---
2.	S. D.	1.44	1.37	1.00
3.	Ave. errors before first success	.749	.785	.705
4.	S. D.	.98	1.08	.84
5.	Ave. errors between first and second success	.361	.350	.315
6.	S. D.	.76	.72	---
7.	Ave. errors before second success	1.11	1.13	1.02
8.	S. D.	1.10	1.01	.93
9.	Ave. successes between errors	.488	.540	---
10.	S. D.	.72	.83	---
11.	Ave; trial of last error	2.18	2.33	3.08
12.	S. D.	2.40	2.47	3.39
13.	Total error runs	.975	.966	1.162
14.	Error runs of length 1	.655	.645	.949
15.	Error runs of length 2	.215	.221	.144
16.	Error runs of length 3	.070	.058	.064
17.	Error runs of length 4	.023	.024	.005
	Autocorrelation of errors			
18.	--one trial apart (c_1)	.479	.486	.288
19.	--two trials apart (c_2)	.310	.292	.195
20.	--three trials apart (c_3)	.201	.187	.127
21.	Alternations of success and failure	1.45	1.43	1.83
22.	(Fail-Fail runs) — (Success-Fail runs)	.000	.020	-.380
23.	Probability of a success following an error	.672	.666	.730

statistics are shown for which the predictions of the linear model have not been worked out (although stat rats could have been run for these).

The results in Table 1 require little comment. Of the 19 possible comparisons between the one-element and linear models, the one-element model comes closer to the data on 17. The greatest differentiation of the models is seen in sequential statistics, lines 13 through 23, and in the trial number of the last failure (lines 11 and 12). The largest absolute discrepancy from data of the one-element predictions occurs with the average trial number of the last failure, but this statistic also has the largest variance of all those considered. Weighing these considerations along with the excellent fits of the one-element model to the data shown in Figs. 2–9, we may conclude that the one-element model provides a more adequate description of these data than does the linear model.

Other paired-associate data favoring the one-element model have been reported in [4]. One dramatic comparison of the two models is provided by considering the expected number of errors (to perfect learning) following an error that occurs on trial n. According to the linear model, the number of errors expected following an error on trial n should be a decreasing function of n, since associative strength is assumed to increase steadily with the number of preceding reinforced trials. In contrast, from the one-element model the expectation is that the average errors following an error on trial n is a constant, $(1 - c)u_1$, which is independent of the trial number on which the error was observed. The point of the matter is that if we observe an error on trial n, then we know the item was not conditioned prior to that trial; hence, we can assume that our learning process "starts" in conditioning state C_0 at the beginning of trial n and that the state of the subject's associative connection has not effectively changed since he started the experiment. We may, so to speak, reset the clock back to the beginning of the experiment for predicting the subject's future behavior on that item.

To get a stable test of these different predictions, the present data from 29 subjects were pooled with the data of 47 other subjects learning 10 paired-associate items under the same conditions except for 14 of the subjects the number of response alternatives was 3, and for 14 there were 8 responses. The varying N's would not affect the constancy or monotone decreasing aspects of the two predictions. For the 760 learning sequences the average number of errors following an error on trial 1, on trial 2, $\cdots$, on trial 6 were calculated. The data beyond trial 6 were not analyzed since the number of cases involved was dropping off rapidly. The results of these calculations are shown in Fig. 10 where the one-element model prediction (i.e., average of all the data points) and a rough approximation to the linear model's predictions are included for comparative purposes. There is little doubt that the one-element prediction is closer to the data, which show remarkable constancy.

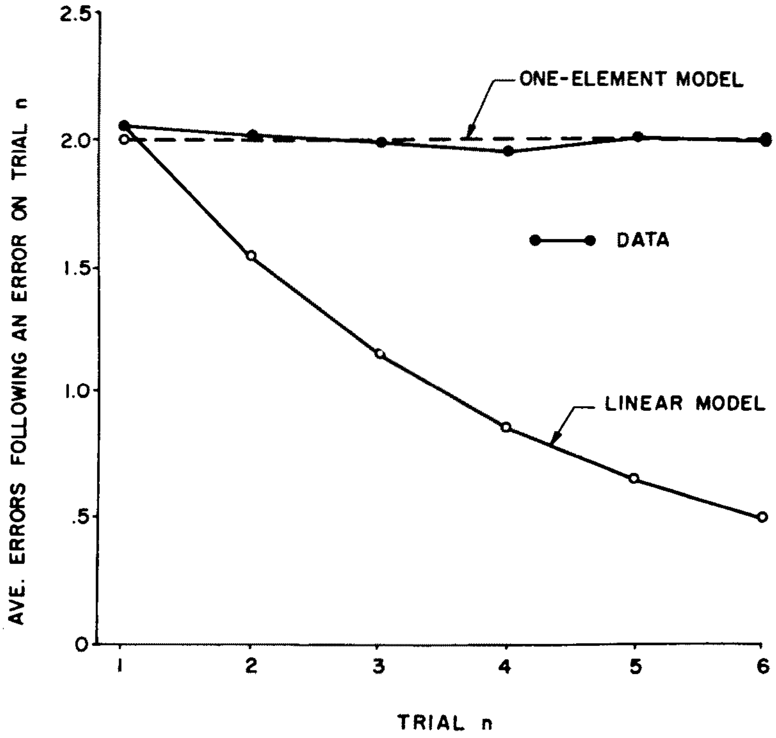

FIGURE 10
Average Number of Errors Following an Error on Trial n
(The one-element and linear predictions are indicated.)

The predicted function for the linear model is not exact since groups with differing N's and c's were pooled in Fig. 10; however, the function does show the relative order of magnitude of differences to be expected from the linear model. The values for the graph were obtained by estimating the average c value across groups (it was .25) and then multiplying successive values on the function by $1 - c$. For example, the average errors observed following an error on trial 1 was 2.05; hence, for trial 2 the linear prediction would be 2.05 (.75) = 1.54, and so on.

Goodness of Fit Considerations

Although the preceding tabulation of various statistics and distributions tells us something about how well the model describes these data, still one legitimately may raise the question of whether there is some summary measure for evaluating the *over-all* goodness of fit of the model to these data. For these purposes a chi-square procedure adapted for stochastic learning models by Suppes and Atkinson [15] from an original paper by Anderson and Goodman [1] may be used. This procedure tests the ability of a model to reproduce the nth order conditional probabilities in the response sequences. Tests involving such quantities have priority in evaluating goodness of fit since the primary reference of stochastic models is to these conditional probabilities. Other statistics of the data (error runs, trial of second success, etc.) are more or less derived from these conditional probabilities and thus should have less priority in deciding over-all goodness of fit. The chi-square test proposed by Anderson and Goodman is most appropriate for those models which assume that the sequence of response random variables is a finite Markov chain (i.e., that current response probability depends upon, say, only one or two prior responses). This happens to be a rather restricted class of learning models; however, the test is practically useful even for chains of infinite order in which current response probability presumably depends upon the entire past history of responses and reinforcements. In practice, such chains can be approximated reasonably well by taking account of only a small number (say 3 or 4) of prior trials when calculating conditional probabilities from the theory.

The chi-square procedure may be illustrated with the present data. The decision was made to look at third-order conditional probabilities using the data from the first six trials of the experiment; beyond trial 6 practically all responses were correct so very little information could be gained by considering the data beyond that point. With two responses (correct and incorrect) there are eight possible sequences of length three. The data were tabulated in an 8×2 table, the entries in each cell corresponding to the frequency with which a given sequence of responses on trials n, $n + 1$, and $n + 2$ was followed by a success (or failure) on trial $n + 3$. For each subject-item sequence, three observations were obtained corresponding to n taking on the values 1, 2, and 3. There were thus $3(290) = 870$ observations in total.

The expected conditional probabilities are readily calculated from the one-element model. For example, four of the eight prior sequences have an error on trial $n + 2$; hence, the expected conditional probability of an error on trial $n + 3$ would be α. The only conditional probability which is troublesome to compute is that of an error given a prior sequence of three successes (with responses prior to trial n being unspecified). This conditional probability is calculated separately for $n = 1, 2, 3$, and then the three results are averaged. Analogous computations from the linear model are extremely simple—in that model the x_n are considered to be statistically independent; hence, one merely averages the response probabilities on trials 4, 5, and 6.

The conditional probabilities calculated above are converted into cell frequencies by multiplying them by the observed frequency of a given prior sequence of three responses (i.e., we multiply by the observed row sums of the table). Chi-square values can then be calculated separately for the one-element model and for the linear model. There are eight rows in the table, each row having one linear constraint (the two entries must sum to the appropriate row total) and for each model we have estimated one parameter (c); hence, each chi square will have seven degrees of freedom. The chi-square values for the observed and expected frequencies were 9.40 for the one-element model and 98.36 for the linear model. Therefore, the test rejects the linear model in its fit to these data but does not reject the one-element model.

Suppes and Chmura [16] have proposed a simple but rigorous procedure for discriminating between the goodness of fit of two models for which the above chi-square values have been calculated. Their statistic, T, is the ratio of the two chi-square values, each divided by its respective degrees of freedom. Under the assumption that one of the models is true, T is distributed as the noncentral F statistic, with a noncentrality parameter equal to the value of an ordinary chi square done on the two sets of expected frequencies (ignoring the data for the moment). For the present case, the value of T is 10.40. This value is so large (an ordinary F table requires only 7.00 for significance at the .01 level) that it would be a mere formality to calculate its exact probability under the assumption that both models fit the data equally well. Hence, we may unequivocally reject the linear model in favor of the one-element model.

Range of Application of the Model

The fact that the one-element model gives an adequate quantitative account of these paired-associate data satisfies one important requisite of a scientific theory, that of being close to the data. If, in addition, the theory is mathematically tractable in that numerous consequences are easily derived in closed form, then indeed we are in a fortunate position. The main task of this paper has been to show that the one-element model is mathematically tractable; those familiar with current work in mathematical learning theory certainly can have no quarrel with this claim. This property of the model is due to the extreme simplicity of its assumptions about the association process. One might effectively argue that the present model nearly achieves the absolute minimum in assumptions for a workable theory of learning.

Once one has demonstrated the predictive validity of a model for a limited class of experimental situations, there remains the task of characterizing more generally those experimental arrangements to which the model may be expected to apply. In the first part of this report, we explicitly restricted the model to the S-R association process and have used simplified experimental situations in which response learning was precluded. Within this restricted domain of PAL, the model has proved extremely useful in investigating the effects on learning of variations in the number of response alternatives and in the reinforcement conditions prevailing during learning [4].

In addition, the model has led us to do experiments in which the guessing probabilities are altered indirectly by varying the proportion of items in the list that have the same correct response (e.g., with 20 items and responses 1 and 2, we have varied the number of items that have 1 as the correct response).

The experimental conditions may differ considerably from those obtaining under paired-associate learning, but still the model may be expected to apply if response learning is precluded. A good example of such an application is to the paradigm that experimenters have called verbal discrimination learning (e.g., Rundquist and Freeman, [14]). In one variant of this experiment, the subject is required to read the correct response from a card on which are printed N alternatives (words, syllables); the subject goes repeatedly through a deck of K such cards until he can give the correct response to all of them. The model has been applied to the results of such an experiment with $N = 2$; its predictive validity proved equally as good as that reported here for the paired-associate task. To cite a further example of work in progress, we are attempting to extend the model to a similar task in which the subject learns to recognize or identify a visual form as one of those that had been shown to him in a "training list" of visual forms.

A further extension of the present work would investigate the modifications in the theory that are required to handle those PAL situations in which the responses per se must be learned. Here again it may prove advantageous to fractionate the problem by utilizing experimental arrangements which primarily involve only response learning. The free verbal recall paradigm [e.g., 2] would appear to serve these purposes. In such experiments the subject is read a number of unrelated words and later is tested for free, unaided recall. With this arrangement, the responses are conditioned presumably to situational and intraverbal cues in a manner analogous to that assumed to occur in PAL response learning. Evidence already exists to indicate that the free verbal recall situation may yield to a simple theoretical analysis. Miller and McGill [12] and Murdock [13] have published quantitative theories which appear to account adequately for their results from free verbal recall experiments. Ultimately, one would like to have a set of combination axioms whereby the assumptions about S-R association and response learning may be combined for predicting results in those experimental situations involving the concurrent operation of these two processes. It may not be presumptuous to suppose that such a development will come about in the next few years.

REFERENCES

[1] Anderson, T. W. and Goodman, L. A. Statistical inference about Markov chains. *Ann. math. Statist.*, 1957, **28**, 89-110.

[2] Bruner, J. S., Miller, G. A., and Zimmerman, C. Discriminative skill and discriminative matching in perceptual recognition. *J. exp. Psychol.*, 1955, **49**, 187-192.

[3] Bower, G. H. Properties of the one-element model as applied to paired-associate

learning. Tech. Rep. No. 31, Psychol. Ser., Inst. for Mathematical Studies in the Social Sciences, Stanford Univ., 1960.

[4] Bower, G. H. A model for response and training variables in paired-associate learning. *Psychol. Rev.*, in press.

[5] Bush, R. R. Sequential properties of linear models. In R. R. Bush and W. K. Estes (Eds.), *Studies in mathematical learning theory.* Stanford: Stanford Univ. Press, 1959. Pp. 215-227.

[6] Bush, R. R. and Mosteller, F. A comparison of eight models. In R. R. Bush and W. K. Estes (Eds.), *Studies in mathematical learning theory.* Stanford: Stanford Univ. Press, 1959. Pp. 293-307.

[7] Bush, R. R. and Sternberg, S. A single-operator model. In R. R. Bush and W. K. Estes (Eds.), *Studies in mathematical learning theory.* Stanford: Stanford Univ. Press, 1959. Pp. 204-214.

[8] Estes, W. K. Toward a statistical theory of learning. *Psychol. Rev.*, 1950, **57**, 94-107.

[9] Estes, W. K. and Burke, C. J. A theory of stimulus variability in learning. *Psychol. Rev.*, 1953, **60**, 276-286.

[10] Estes, W. K. Component and pattern models with Markovian interpretations. In R. R. Bush and W. K. Estes (Eds.), *Studies in mathematical learning theory.* Stanford: Stanford Univ. Press, 1959. Pp. 9-52.

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

[12] Miller, G. A. and McGill, W. J. A statistical description of verbal learning. *Psychometrika*, 1952, **17**, 369-396.

[13] Murdock, B. B. The immediate retention of unrelated words. *J. exp. Psychol.*, 1960, **60**, 222-234.

[14] Rundquist, W. N. and Freeman, M. Roles of association value and syllable familiarization in verbal discrimination learning. *J. exp. Psychol.*, 1960, **59**, 396-401.

[15] Suppes, P. and Atkinson, R. C. *Markov learning models for multi-person interactions.* Stanford: Stanford Univ. Press, 1960.

[16] Suppes, P. and Chmura, H. A statistical test for comparative goodness of fit of alternative learning models. Tech. Rep. No. 36, Psychol. Ser., Inst. for Mathematical Studies in the Social Sciences, Stanford Univ., 1961.

Manuscript received 9/1/60

Revised manuscript received 1/11/61