JOURNAL OF MATHEMATICAL PSYCHOLOGY: 1, 285-315 (1964)

A Comparison of Paired-Associate Learning Models Having Different Acquisition and Retention Axioms[1]

RICHARD C. ATKINSON AND EDWARD J. CROTHERS

Stanford University, Stanford, California

Several alternative interpretations of all-or-none processes for paired-associate learning and concept formation are examined. These models, along with three linear models, are applied to data from eight paired-associate learning experiments. The principal analyses involve goodness-of-fit tests for observed response sequences and conditional probabilities. The results favor a three-process model that postulates a distinction between long-term and short-term retention and allows for forgetting between successive presentations of the same stimulus item.

I. INTRODUCTION

In recent articles Bower (1961, 1962), Crothers (1962), Estes (1960, 1961), Suppes and Ginsberg (1963), and others 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 principal 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 by guessing. (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 $\bar{C}$ the element is unconditioned. The element starts out in state $\bar{C}$ and subsequently moves to state C as specified by the transition matrix

$$\begin{array}{c} \\ C \\ \bar{C} \end{array} \begin{array}{c} \begin{array}{cc} C & \bar{C} \end{array} \\ \begin{bmatrix} 1 & 0 \\ c & 1-c \end{bmatrix} \end{array}. \tag{1}$$

[1] Support for this research was provided by the National Science Foundation (Grant G-24264) and by the U. S. Office of Education (Contract OE 3-14-020). Computer time was paid for by the National Institute of Mental Health (Grant 6154-02).

By and large, the results reported by Bower, Crothers, Estes, and Suppes and Ginsberg indicate a remarkably close correspondence between observed and predicted values for the one-element model. The agreement 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 nonstationary 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 nonstationary 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

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

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. However, there are at least two reasons why the two-element model is unsatisfactory for paired-associate learning. First, while it is reasonable to equate the parameter g with the reciprocal of the number of response alternatives, there seems to be no convincing experimental interpretation of a value of g' estimated from data. Secondly, we shall see that even when g' is estimated from data, certain predictions of the two-element model are inaccurate.

The aim of this paper is to develop a model that is conceptually quite different from the two-element model, but which predicts the nonstationarity effect and is relatively more accurate otherwise. We cite paired-associate data using the anticipation method in comparing the goodness-of-fit of the proposed model with the fits of the one-element and the two-element models. Also, for purposes of comparison, we examine several linear models. Thus the models compared include most of those previously proposed to account for paired-associate learning, as well as a variety of new formulations.

Because of the particular data to be analyzed here, all of the models will be developed 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 so we assume that the probability of a correct response by guessing is $1/r$. On each trial the stimuli are exhibited singly in a new random order. When a stimulus is presented the subject is required to make a response and is then informed of the correct response.

To introduce the proposed model, let us sketch the main argument that motivated us toward this approach. When a set of models collectively fails to provide accurate predictions of response sequences, a major reason would seem to be that some psychological process not represented in the models is influencing behavior. A prime candidate for such a process would appear to be the occurrence of forgetting between successive presentations of the same item; certainly, forgetting is as ubiquitous as acquisition. Appreciable forgetting of individual consonant syllables and paired associates over short intervals of time has become an established fact (Melton, 1936b; Murdock, 1961; Peterson and Peterson, 1959); in these experiments, the subject counted backwards during the interval between the reinforced presentation and the test. Other experiments by Peterson, Saltzman, Hillner, and Land (1962) resembled conventional paired-associate learning studies more closely in that the predetermined reinforcement-test intervals were occupied with presentations of other items. The empirical findings in these studies enhance the general appeal of the suggestion that a forgetting mechanism can increase the accuracy of paired-associate models. Of course, there remains the problem of exactly how to express such a process in a model for experiments where the subject learns to a criterion.

To specify a learning model with forgetting, we begin by noting that forgetting has a natural Markovian interpretation as a transition to a lower state of learning. Since the subject eventually learns to criterion, it seems important to introduce the distinction between long-term retention and short-term retention. In the latter state, forgetting can occur and corresponds to regression to a state in which errors are possible.[2] Beyond these general remarks, there are a variety of ways in which one can pursue the mathematical formulation of learning models that incorporate a forgetting process. In the next section we shall consider only the model that appears most promising. After the data have been presented, it will be easier to see why certain alternative models embodying forgetting processes are less satisfactory.

II. A LEARNING MODEL WITH ENCODING AND FORGETTING AXIOMS

The model assumes four stages of learning: L, S, F, and U. Learning is postulated to consist of encoding the stimulus (Lawrence, 1963) followed by associating the enco-

[2] A distinction in terms of the retention interval (Melton, 1963a) attributes forgetting to both processes and is less convenient for our immediate purposes.

ded stimulus to the correct response.[3] Before encoding has occurred, the stimulus is said to be in state U (uncoded); in this state the subject is assumed to respond by guessing randomly among the r alternatives. After the stimulus is encoded, it can become associated to the correct response. Once the association forms the stimulus element is absorbed in state L (long-term memory) and the subject makes no errors on subsequent presentations of the item. Transitions between the intermediate states S and F represent events assumed to intervene between the encoding and association phases. State S is a short-term memory state, expressing the notion that a temporary connection between the encoded stimulus and the response may form prior to establishing the permanent association; while the association is temporarily stored the correct response occurs with probability 1. However, the temporary connection is susceptible to forgetting, in which case the stimulus element is said to pass into state F. Here, as in state U, the subject guesses randomly; however, forgetting is only partial, since the encoding is retained.

Stated more precisely, encoding for a given stimulus item occurs at most on one trial; the probability that encoding occurs on trial n given that it has not occurred on previous trial is c. If an item is presented that has already been encoded (either on the present trial or on an earlier trial), then with probability a it goes into state L and with probability 1-a it goes into state S. Thus, after each presentation, an encoded item is in either state L or S, and if the item were to be presented again immediately the subject would make the correct response with probability 1. However, other events intervene from one presentation of an item to its next presentation, and during this period we assume there is a probability f that an item in state S will move back to state F. We assume the value of f depends upon the number and type of intervening items; also, f depends upon the exposure time of the given item, for this affects the repetition rate and hence the slope of the forgetting function (Hellyer, 1962; Peterson *et al.*, 1962).

Given the above assumptions, it can be shown that moves among the four states are described by the following transition matrix and response probability vector:

$$\begin{array}{c} \\ L \\ S \\ F \\ U \end{array}\!\!\begin{array}{c} \begin{array}{cccc} L & S & F & U \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ a & (1-a)(1-f) & (1-a)f & 0 \\ a & (1-a)(1-f) & (1-a)f & 0 \\ ca & c(1-a)(1-f) & c(1-a)f & 1-c \end{bmatrix} \end{array} \quad \begin{array}{c} \text{Pr (correct)} \\ \begin{bmatrix} 1 \\ 1 \\ g \\ g \end{bmatrix} \end{array} \tag{3}$$

where $g = 1/r$; throughout the paper we shall use g to denote the guessing probability. Before proceeding with the derivations, let us mention a few features of this model. First, it is clear that the predicted probability of a correct response can increase over

[3] As viewed here, the encoding process is no more than a heuristically useful component of the model. There seems to be no point in endowing it with special psychological properties, for reasons which will become apparent later.

trials prior to the last error. This is because errors later in learning are more likely to have occurred in state F than are earlier errors; hence, the later errors would have been more frequently preceded by runs of errorless trials in state S. Another property which seems desirable is that the model is qualitatively in accord with overlearning phenomena; postcriterion training trials produce transitions from S to L, thereby increasing retention.

III. LINEAR MODELS

For convenience, we classify under this heading all learning models that assume at least one of the factors governing the trial by trial change in response probability is a linear process. The simplest such model is the single-operator linear model (Bush and Mosteller, 1955; Bush and Sternberg, 1959). This model assumes that the probability of the reinforced response increases according to the equation

$$p_{n+1} = (1 - \theta)\, p_n + \theta \tag{4}$$

where $p_1 = 1/r$. Modifications of the above axiom have frequently been applied to probability discrimination learning as well as to paired-associate learning.

Recently, the ability of the single-operator model to account for paired-associate learning has been questioned. The model has been compared unfavorably with the one-element model (Bower, 1961; Estes, 1961; Estes, Hopkins, and Crothers, 1960), touching off a controversy between proponents of all-or-none learning and incremental learning. Our aim in this article is not to support either theoretical position. Instead, we assess only the relative merits of the particular models presented here.

In addition to the single-operator linear model, we shall examine two other models which include linearity assumptions. Since these models are more complex than the original linear model and contain two parameters, they are especially useful in providing comparisons with Markovian processes having more than one parameter. The first linear model (Norman, 1963) assumes a two-phase learning process. An event called "first-learning" is postulated that occurs on at most one trial for any stimulus item; the probability that first-learning occurs on trial k given that it has not occurred on a previous trial is c. A subject's probability of making a correct response depends on the trial of first-learning. Specifically the probability of a correct response on trial n given that first-learning occurred on trial k is

$$p_n = \begin{cases} g, & \text{for} \quad n \leqslant k \\ 1 - (1 - g)\,(1 - \theta)^{n-k}, & \text{for} \quad n > k. \end{cases} \tag{5}$$

Thus, for k trials (where k is geometrically distributed with parameter c) no learning occurs, whereas after trial k a linear learning process takes over of the form specified by Eq. 4. Note that Norman's two-phase model reduces to the one-element model

when $\theta = 1$, and to the simple linear model of Eq. 4 when $c = 1$. The reader is referred to Norman's paper for a fuller discussion of the properties and interpretation of the model.

An alternative two-parameter model that incorporates a linear learning process also has been developed by Norman (1964). In this model, the probability of a correct response on trial $n + 1$ is given by the following equation:

$$p_{n+1} = \begin{cases} (1 - \theta)\, p_n + \theta, & \text{with probability } c \\ p_n, & \text{with probability } 1 - c. \end{cases} \tag{6}$$

Thus on each trial exactly one of two events can occur. With probability $1 - c$ no learning takes place, or with probability c the response probability receives an increment described by the linear transformation given in Eq. 4. Once again, if $\theta = 1$ this process reduces to the one-element model, whereas if $c = 1$ we have the simple linear model. Using Norman's terminology, we shall refer to this combination of the all-or-none and linear axioms as the random-trial-incremental model; henceforth, abbreviated as the RTI model.

IV. PREDICTIONS FOR THE LONG-SHORT MODEL

We now derive a few basic predictions for the model described in Section II; henceforth, for simplicity we shall refer to this model as the LS model, a designation that emphasizes the role of the long-term and short-term retention states. We present those predictions that are particularly helpful in making comparisons among the various models discussed so far. The derivations are carried out for a single stimulus item because later, when we analyze data, it is assumed that the stimulus items are stochastically independent and identical. Throughout the paper we let U_n, F_n, S_n, and L_n denote the events of being in state U, F, S, and L respectively at the start of trial n; also e_n and c_n denote the occurrence of an error and of a correct response on trial n. Further, u_n, f_n, and s_n are used to denote the probabilities of events U_n, F_n, and S_n, respectively.

Learning Curve

For brevity, let $t_n = f_n + s_n$. Then, from the matrix in Eq. 3 we obtain

$$u_n = (1 - c)^{n-1} \tag{7a}$$

$$s_n = (1 - a)(1 - f)\, t_{n-1} + c(1 - a)(1 - f)\, u_{n-1} \tag{7b}$$

$$f_n = f(1 - a)\, t_{n-1} + cf(1 - a)\, u_{n-1}. \tag{7c}$$

Adding Eqs. 7b and 7c yields

$$t_n = (1 - a)\, t_{n-1} + c(1 - a)\,(1 - c)^{n-2}.$$

The solution of this difference equation is (cf. Atkinson and Estes, 1963, p. 148)

$$t_n = (1 - a)^{n-1}\, t_1 + c(1 - a)^{n-1} \sum_{i=0}^{n-2} \left(\frac{1 - c}{1 - a}\right)^i.$$

Or, since we asume $t_1 = 0$,

$$t_n = c(1 - a)^{n-1} \sum_{i=0}^{n-2} \left(\frac{1 - c}{1 - a}\right)^i.$$

There are two cases to be distinguished:

$$t_n = c(n - 1)\,(1 - a)^{n-1}, \qquad \text{for} \qquad c = a \tag{8a}$$

and

$$t_n = \frac{c(1 - a)}{c - a}\,[(1 - a)^{n-1} - (1 - c)^{n-1}], \qquad \text{for} \qquad c \neq a. \tag{8b}$$

Then,

$$f_n = f t_n = \frac{fc(1 - a)}{c - a}\,[(1 - a)^{n-1} - (1 - c)^{n-1}] \tag{9a}$$

$$s_n = (1 - f)\, t_n = \frac{c(1 - f)\,(1 - a)}{c - a}\,[(1 - a)^{n-1} - (1 - c)^{n-1}] \tag{9b}$$

for $c \neq a$. When $c = a$, the expressions for f_n and s_n are obvious from Eq. 8a.

Since errors occur with probability $1 - g$ in either state U or F, the probability of an error on trial n is

$$\begin{aligned} \Pr(e_n) &= (1 - g)\,(u_n + f_n) \\ &= (1 - g)\left\{(1 - c)^{n-1} + \frac{fc(1 - a)}{c - a}\,[(1 - a)^{n-1} - (1 - c)^{n-1}]\right\} \end{aligned} \tag{10}$$

E(*T*), Expected Total Errors Per Item

This quantity is the sum of the expected total errors in state U and in state F, which we denote as $E(U)$ and $E(F)$, respectively. It is well known that

$$E(U) = \frac{1 - g}{c}.$$

To find $E(F)$ we begin by deriving the probability that the subject eventually enters state F. First, define ρ as the probability of eventually returning from state S to state F; it is easily shown that

$$\rho = \frac{(1-a)f}{a+(1-a)f}.$$

The probability that the subject eventually enters state F is simply

$$\begin{aligned} &c(1-a)f + (1-c)\,c(1-a)f + (1-c)^2\,c(1-a)f + \cdots \\ &\quad + c(1-a)\,(1-f)\,\rho + (1-c)\,c(1-a)\,(1-f)\,\rho \\ &\quad + (1-c)^2\,c(1-a)\,(1-f)\,\rho + \cdots \\ &\quad = (1-a)\,[f + (1-f)\,\rho] = \rho. \end{aligned}$$

Given that the subject has entered state F, the equation for the expected number of errors in state F has been found previously (Crothers, 1963, p. 5) and is as follows:

$$\frac{(1-g)\,[a+(1-a)f]}{a}.$$

Hence

$$E(F) = \frac{(1-g)\,\rho[a+(1-a)f]}{a} = \frac{(1-g)\,(1-a)f}{a}.$$

Combining these results we obtain

$$\begin{aligned} E(T) &= E(U) + E(F) \\ &= (1-g)\left[\frac{1}{c} + \frac{(1-a)f}{a}\right]. \end{aligned} \tag{11}$$

Of course this expression could have been computed directly from Eq. 10; however, the derivation was carried out in this way because some of the intermediate results will be needed later.

Distribution of the Trial Number of the Last Error

Let v_n be the probability that the last error occurs on trial n. Further, let b_U denote the probability of the event "no further errors after a response in state U", and likewise define b_F for state F. Then

$$\begin{aligned} v_n &= \Pr(e_n \cap U_n)\,b_U + \Pr(e_n \cap F_n)\,b_F \\ &= (1-g)\,[u_n b_U + f_n b_F]. \end{aligned}$$

Further,

$$b_F = a + (1 - a)(1 - f)(1 - \rho + g\rho b_F) + gf(1 - a)\, b_F .$$

This equation was obtained by considering all of the ways in which the event "no further errors after a response in state F" can occur. For example, the term $(1 - a)(1 - f)\, g\rho b_F$ represents the probability of the joint event "pass to state S, eventually return to state F, make a correct response on the next trial, and make no further errors." Algebraic operations yield

$$b_F = \frac{w}{z} \tag{12}$$

where

$$w = a + (1 - a)(1 - f)(1 - \rho)$$

$$z = 1 - (1 - a)(\rho - \rho f + f)\, g.$$

Likewise,

$$b_U = ca + c(1 - a)(1 - f)(1 - \rho + g\rho b_F) + c(1 - a)\, gfb_F + g(1 - c)\, b_U .$$

Simplifying and substituting for b_F from Eq. 12 yields

$$b_U = \frac{cw}{z[1 - (1 - c)\, g]} . \tag{13}$$

By using Eqs. 12 and 13 in the above expression for v_n we obtain

$$v_n = \frac{w}{z}(1 - g)\left[\frac{cu_n}{1 - g(1 - c)} + f_n\right], \tag{14}$$

where u_n and f_n are given by Eqs. 7a and 9a, respectively.

For the probability that the subject makes no errors, which is denoted as v_0, we have

$$v_0 = gb_U .$$

This completes the derivation.

THE PROBABILITY OF AN ERROR ON TRIAL $n + 1$, CONDITIONAL ON AN ERROR ON TRIAL n

We begin by finding the probability of an error on both trial n and $n + 1$; namely,

$$\begin{aligned} \Pr(e_n \cap e_{n+1}) &= \Pr(e_n \cap e_{n+1} \cap U_n) + \Pr(e_n \cap e_{n+1} \cap F_n) \\ &= (1 - g)^2 \{[c(1 - a)f + 1 - c]\, u_n + f(1 - a) f_n\}. \end{aligned} \tag{15}$$

Then, of course

$$\Pr(e_{n+1} \mid e_n) = \frac{\Pr(e_{n+1} \cap e_n)}{(1 - g)(u_n + f_n)} . \tag{16}$$

To study the behavior of $\Pr(e_{n+1} \mid e_n)$ as n increases, we substitute Eq. 15 in Eq. 16 and divide both numerator and denominator by u_n:

$$\Pr(e_{n+1} \mid e_n) = \frac{(1-g)\{[c(1-a)f + 1 - c] + f(1-a)(f_n/u_n)\}}{1 + (f_n/u_n)}. \tag{17}$$

Regarded as a function of n, the above equation is dominated by the ratio f_n/u_n in the denominator, and from Eqs. 7a and 9a (for $a \neq c$)

$$\frac{f_n}{u_n} = \frac{c(1-a)f}{c-a}\left[\left(\frac{1-a}{1-c}\right)^{n-1} - 1\right].$$

We see that the quantity $[(1-a)/(1-c)]^{n-1}$ increases as n increases if, and only if, a is less than c. Under this condition, therefore, $\Pr(e_{n+1} \mid e_n)$ decreases with increasing values of n.

Probabilities of Response Sequences over Trials n to $n+3$

In this section we present predictions for four-tuple response sequences; these quantities are particularly useful in making comparisons among the various models described in this paper. For simplicity we denote the 16 possible outcome sequences over trials n to $n+3$ as follows:

$$\begin{aligned}
O_{1,n} &= c_n c_{n+1} c_{n+2} c_{n+3} & O_{9,n} &= e_n c_{n+1} c_{n+2} c_{n+3} \\
O_{2,n} &= c_n c_{n+1} c_{n+2} e_{n+3} & O_{10,n} &= e_n c_{n+1} c_{n+2} e_{n+3} \\
O_{3,n} &= c_n c_{n+1} e_{n+2} c_{n+3} & O_{11,n} &= e_n c_{n+1} e_{n+2} c_{n+3} \\
O_{4,n} &= c_n c_{n+1} e_{n+2} e_{n+3} & O_{12,n} &= e_n c_{n+1} e_{n+2} e_{n+3} \\
O_{5,n} &= c_n e_{n+1} c_{n+2} c_{n+3} & O_{13,n} &= e_n e_{n+1} c_{n+2} c_{n+3} \\
O_{6,n} &= c_n e_{n+1} c_{n+2} e_{n+3} & O_{14,n} &= e_n e_{n+1} c_{n+2} e_{n+3} \\
O_{7,n} &= c_n e_{n+1} e_{n+2} c_{n+3} & O_{15,n} &= e_n e_{n+1} e_{n+2} c_{n+3} \\
O_{8,n} &= c_n e_{n+1} e_{n+2} e_{n+3} & O_{16,n} &= e_n e_{n+1} e_{n+2} e_{n+3}.
\end{aligned} \tag{18}$$

These designations will be used throughout this paper. Although this usage may seem inconvenient, it greatly reduces the complexity of subsequent expressions.

From the model one can derive expressions for $\Pr(O_{i,n})$ and from these an array of other quantities can be computed. For example,

$$\begin{aligned}
\Pr(c_n e_{n+2}) &= \Pr(O_{3,n}) + \Pr(O_{4,n}) + \Pr(O_{7,n}) + \Pr(O_{8,n}), \\
\Pr(e_n c_{n+3}) &= \Pr(O_{9,n}) + \Pr(O_{11,n}) + \Pr(O_{13,n}) + \Pr(O_{15,n}),
\end{aligned}$$

and so forth.

We will not present the derivations for $\Pr(O_{i,n})$ here, since they are straightforward and involve only elementary probability theory. (Readers not familiar with the methods involved in such derivations can consult Atkinson and Estes (1963).) However, the

derivations are lengthy and consequently it is of value to present the full array of predictions. They are as follows:

$$
\begin{aligned}
&\Pr(O_{1,n}) = (1 - s_n - f_n - u_n) + (s_n + gf_n)(a + xA_1) + gu_n[c(a + xA_1) + gB_1] \\
&\Pr(O_{2,n}) = (s_n + gf_n)xA_2 + gu_n[cxA_2 + gB_2] \\
&\Pr(O_{3,n}) = (s_n + gf_n)xA_3 + gu_n[cxA_3 + gB_3] \\
&\Pr(O_{4,n}) = (s_n + gf_n)xA_4 + gu_n[cxA_4 + gB_4] \\
&\Pr(O_{5,n}) = (s_n + gf_n)yA_1 + gu_n[cyA_1 + (1 - g)B_1] \\
&\Pr(O_{6,n}) = (s_n + gf_n)yA_2 + gu_n[cyA_2 + (1 - g)B_2] \\
&\Pr(O_{7,n}) = (s_n + gf_n)yA_3 + gu_n[cyA_3 + (1 - g)B_3] \\
&\Pr(O_{8,n}) = (s_n + gf_n)yA_4 + gu_n[cyA_4 + (1 - g)B_4] \\
&\Pr(O_{9,n}) = (1 - g)f_n(a + xA_1) + (1 - g)u_n[c(a + xA_1) + gB_1] \\
&\Pr(O_{10,n}) = (1 - g)f_nxA_2 + (1 - g)u_n[cxA_2 + gB_2] \\
&\Pr(O_{11,n}) = (1 - g)f_nxA_3 + (1 - g)u_n[cxA_3 + gB_3] \\
&\Pr(O_{12,n}) = (1 - g)f_nxA_4 + (1 - g)u_n[cxA_4 + gB_4] \\
&\Pr(O_{13,n}) = (1 - g)f_nyA_1 + (1 - g)u_n[cyA_1 + (1 - g)B_1] \\
&\Pr(O_{14,n}) = (1 - g)f_nyA_2 + (1 - g)u_n[cyA_2 + (1 - g)B_2] \\
&\Pr(O_{15,n}) = (1 - g)f_nyA_3 + (1 - g)u_n[cyA_3 + (1 - g)B_3] \\
&\Pr(O_{16,n}) = (1 - g)f_nyA_4 + (1 - g)u_n[cyA_4 + (1 - g)B_4]
\end{aligned}
\tag{19}
$$

where

$$
\begin{aligned}
x &= (1 - a)(1 - f + fg), \\
y &= (1 - a)(1 - g)f.
\end{aligned}
$$

And

$$
\begin{aligned}
A_1 &= a + x(1 - y), & A_3 &= y(1 - y), \\
A_2 &= xy, & A_4 &= y^2, \\
B_1 &= (1 - c)\{ac + cx(1 - y) + g(1 - c)[c(1 - y) + g(1 - c)]\}, \\
B_2 &= (1 - c)\{cxy + g(1 - c)[1 - c(1 - y) - g(1 - c)]\}, \\
B_3 &= (1 - c)\{cy(1 - y) + (1 - g)(1 - c)[c(1 - y) + g(1 - c)]\}, \\
B_4 &= (1 - c)\{cy^2 + (1 - g)(1 - c)[1 - c(1 - y) - g(1 - c)]\}.
\end{aligned}
$$

In order to make predictions from Eq. 19, estimates of the parameters a, f, and c are needed. There are many ways of making these estimates, but one simple method is to minimize the χ^2 associated with the O_i events.[4] To illustrate the method let $\Pr(O_{i,n}; a, f, c)$ denote the probability of the event $O_{i,n}$, where a, f, and c have been listed to make explicit the fact that the expression is a function of the three parameters. Further, let $N(O_{i,n})$ denote the observed frequency of stimulus items that display outcome O_i over trials n to $n + 3$, and let

$$T = N(O_{1,n}) + N(O_{2,n}) + \cdots + N(O_{16,n}).$$

[4] For a discussion of the χ^2 minimum method of estimation see Cramér (1951, pp. 424-441).

Then we define the function

$$\chi^2(a, f, c) = \sum_{i=1}^{16} \frac{[T \Pr(O_{i,n}; a, f, c) - N(O_{i,n})]^2}{T \Pr(O_{i,n}; a, f, c)} \tag{20}$$

and select our estimates of a, f, and c 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 carry out a systematic search of possible parameter values until a minimum χ^2 is obtained that is accurate to three decimal places.[5] If we assume that all the stimulus items are stochastically independent and identical, then under the null hypothesis it can be shown that this minimum χ^2 has the usual limiting distribution with 12 degrees of freedom.

The minimum χ^2 has several desirable properties as an estimation procedure; the resulting estimates are consistent (as the sample size increases the estimates converge stochastically to the parameter value), and asymptotically efficient (as the sample size increases the variance of the estimates approach the minimal variance attainable for any consistent estimate of the parameter, and the distribution of the estimate approaches normality). The minimum χ^2 also provides a measure of the adequacy of any single model and, if the degrees of freedom are equal, a method for directly comparing the fit of several models. If several models are being analyzed, each involving a different number of free parameters, then the probability levels of the χ^2's may be compared. The degrees of freedom associated with a model that requires k parameters to be estimated from the data are

$$df = 16 - k - 1.$$

In the above equation one degree of freedom has been subtracted because of the restriction that the 16 probabilities must sum to 1.

V. DATA ANALYSES

Descriptions of Experiments

In this section we analyze data from eight paired-associate learning experiments that all utilize the same general experimental procedure. At the start of an experiment the subject is told the responses available to him; each alternative occurs equally often as the to-be-learned response and hence the probability of a correct response by guessing is roughly $1/r$ (where r is the number of alternative responses). A response is obtained from the subject on each presentation of an item and he is informed of the correct answer following his response.

[5] The computer program for the search procedure is available through the Institute for Mathematical Studies in the Social Sciences, Stanford University, Stanford, California.

TABLE 1

FEATURES OF THE EXPERIMENTAL PROCEDURE

Experiment	Number of stimuli	Number of responses	Number of subjects	Pr (c_5)
Ia	9	3	26	.95
Ib	18	3	16	.91
II	12	3	65	.83
III	12	4	40	.75
IV	16	4	20	.84
Va	12	4	40	.60
Vc	12	4	40	.71
Ve	12	4	40	.85

Relevant details of each experiment are given in Table 1. Experiments Ia and Ib were run with college students. For both experiments the stimuli were Greek letters and the responses were the low association trigrams RIX, FUB, and GED; the experiments differed in that one used a 9 item stimulus list and other an 18 item list. Experiment II was also run with college students using 12 Greek letters as stimuli and the numbers 4, 5, and 6 as responses. Experiment III was run with 3rd and 4th grade students using 12 Greek letters as stimuli and the numbers 2, 3, 4, and 5 as responses. Experiment IV was run with college students using double digit numbers as stimuli and the letters *A*, *B*, *C*, and *D* as responses. For Experiments I-IV, the experimental procedure (method of stimulus display, presentation rate, etc.) was the same as described by Bower (1961). In Experiment V, a group of four and five year old children learned a list of paired-associates each day for five consecutive days. The lists were composed of double digit numbers as stimuli and letters as responses but the stimuli and responses were different for each list. To simplify the discussion, only results for days 1, 3, and 5 are presented (labeled Experiments Va, Vc, and Ve respectively); however, these data are representative of the results for the full experiment. A complete description of the experimental procedure and results is available elsewhere (Hansen, 1963).

ANALYSIS OF THE FOUR-TUPLE DATA

We now turn to an analysis of the response tuples described by Eq. 18 for trials 2 to 5. For the experiments discussed in this paper, these statistics are of particular importance because a major portion of the learning occurred during the first five trials. This fact is indicated in the last column of Table 1 where Pr (c_5) is presented; in five of the eight experiments the subjects have reached a correct response level of 0.83 or better on trial 5.

TABLE 2

OBSERVED FREQUENCIES FOR THE $O_{i,2}$ EVENTS

	Experiment							
	Ia	Ib	II	III	IV	Va	Vc	Ve
$N(O_{1,2})$	123	125	303	160	117	82	144	216
$N(O_{2,2})$	3	3	14	13	3	11	18	4
$N(O_{3,2})$	6	10	19	16	10	14	23	17
$N(O_{4,2})$	1	4	12	11	1	13	9	6
$N(O_{5,2})$	16	21	54	24	15	22	28	34
$N(O_{6,2})$	3	0	17	6	3	21	14	16
$N(O_{7,2})$	5	6	32	18	9	20	12	12
$N(O_{8,2})$	2	3	18	7	6	31	13	12
$N(O_{9,2})$	43	55	125	57	54	58	62	66
$N(O_{10,2})$	1	5	15	9	7	13	14	4
$N(O_{11,2})$	7	10	25	27	9	34	25	17
$N(O_{12,2})$	2	2	17	14	10	18	14	7
$N(O_{13,2})$	15	30	61	33	34	34	28	29
$N(O_{14,2})$	0	1	19	25	8	21	20	8
$N(O_{15,2})$	6	6	30	24	22	26	21	19
$N(O_{16,2})$	1	7	19	36	12	62	35	13
T	234	288	780	480	320	480	480	480

Table 2 presents the observed frequencies of the $O_{i,2}$ events for each study. Experilent Ia has 26 subjects each run on a list of 9 stimulus items, and hence there are $26 \times 9 = 234$ item-response sequences. As indicated in the table, for 123 sequences no errors occurred on trials 2, 3, 4, and 5; 3 sequences displayed no errors on trials 2, 3, and 4 but an error on trial 5, and so on.

The χ^2 minimization procedure described by Eq. 20 was applied to the data given in Table 2 for each of the paired-associate models. Table 3 presents the parameter estimates associated with the minimum χ^2 values. For the LS model the minimization was carried out for the general case (where the three parameters a, f, and c were estimated simultaneously), and also for the special case where $c = 1$; henceforth, we shall refer to the first case as the LS-3 model and the second case as the LS-2 model (the 3 and 2 designate the number of free parameters to be estimated). In five of the eight experiments the estimate of c for the LS-3 model was virtually equal to 1; hence, in these instances the LS-3 model reduced to the two parameter version.

One property of parameter estimates that appears desirable is monotonicity over the three sets of Experiment V data. This property seems reasonable since the subjects

TABLE 3

PARAMETER ESTIMATES FOR THE VARIOUS MODELS

Model	Parameter	Experiment							
		Ia	Ib	II	III	IV	Va	Vc	Ve
One-element	c	.383	.328	.273	.203	.281	.125	.172	.289
Linear	θ	.414	.328	.289	.258	.297	.164	.250	.336
Two-phase	c	.563	.484	.352	.359	.398	.227	.406	.422
	θ	.664	.633	.695	.563	.648	.500	.477	.656
RTI	c	.531	.461	.344	.328	.367	.219	.359	.438
	θ	.820	.805	.867	.797	.859	.727	.711	.789
LS-2	a	.352	.305	.250	.188	.266	.109	.156	.258
	f	.719	.805	.805	.789	.836	.844	.727	.680
LS-3	a	.367	.352	.250	.188	.289	.109	.156	.266
	f	.648	.375	.805	.789	.789	.844	.727	.688
	c	.844	.500	1.000	1.000	.789	1.000	1.000	.992
Two-element	g'	.883	.852	.922	.891	.922	.797	.859	.844
	b	.391	.398	.227	.078	.195	.133	.016	.227
	a	.539	.477	.344	.320	.359	.219	.352	.477

and procedures were the same, and the over-all proportion of errors decreased steadily over the five experimental sessions. However, Table 3 reveals monotonicity only for the parameter estimates associated with the LS-3 and LS-2 models (and, of course, for the one-parameter models).

For each of the models, the ranks of the magnitude of the parameter estimates were consistent over the eight experiments. For the two-phase model and the RTI model the estimate of c was consistently less than that of θ; for the LS model, $\hat{a} < \hat{f} < \hat{c}$; and for the two-element model $\hat{b} < \hat{a} < \hat{g}'$. It is interesting to note that in the RTI model $\hat{\theta} \geqslant 0.71$, and in the two-element model $\hat{g}' \geqslant 0.79$. These high estimates imply that for both models the first stepwise increment in response probability is rather large.

As indicated earlier, Experiments Ia and Ib are comparable except that the former study used a list of 9 items and the latter an 18-item list. In regard to the LS-3 model, it is interesting to note that the conditioning parameter a is about the same for both list lengths. However, the list-length variable not only influences $\hat{f}$, but $\hat{c}$ as well.

TABLE 4

MINIMUM χ^2 VALUES

Experiment	One-element	Linear model	Two-phase	RTI	LS-2	LS-3	Two-element
Ia	30.30	50.92	17.51[a]	9.74[a]	6.75[a]	5.67[a]	9.30[a]
Ib	39.31	95.86	18.25[a]	13.09[a]	19.69[a]	12.42[a]	12.74[a]
II	62.13	251.30	54.78	29.11	3.73[a]	3.73[a]	28.46
III	150.66	296.30	95.44	51.12	33.02	33.02	47.13
IV	44.48	146.95	22.39[a]	10.66[a]	12.32[a]	10.77[a]	10.32[a]
Va	102.02	201.98	59.20	40.17	24.41[a]	24.41[a]	39.47
Vc	246.96	236.15	99.97	46.43	27.12[a]	27.12	34.75
Ve	161.03	262.56	126.05	84.07	20.12[a]	20.12[a]	77.39
Total χ^2	836.89	1542.02	493.59	284.39	147.16	137.26	259.56
df	14	14	13	13	13	12	12

[a] Not significant at .01 level.

TABLE 5

OBSERVED AND PREDICTED RESPONSE SEQUENCE PROPORTIONS FOR EXPERIMENT II

Outcomes	Observed proportion	One-element	Linear model	Two-phase	RTI	Long-short	Two-element
O_1	.389	.362	.220	.328	.354	.390	.357
O_2	.018	.007	.045	.008	.017	.017	.018
O_3	.024	.015	.069	.022	.028	.029	.029
O_4	.015	.014	.014	.010	.011	.020	.011
O_5	.069	.047	.112	.066	.063	.064	.062
O_6	.022	.014	.023	.012	.013	.020	.013
O_7	.041	.029	.035	.028	.026	.034	.026
O_8	.023	.028	.007	.021	.020	.023	.020
O_9	.161	.178	.198	.210	.189	.164	.188
O_{10}	.019	.014	.041	.014	.018	.020	.018
O_{11}	.032	.029	.062	.035	.034	.034	.034
O_{12}	.022	.028	.013	.021	.020	.023	.020
O_{13}	.079	.093	.101	.102	.092	.074	.091
O_{14}	.024	.028	.021	.024	.024	.023	.024
O_{15}	.038	.059	.032	.055	.051	.039	.050
O_{16}	.024	.055	.007	.042	.040	.026	.039

Table 4 presents the minimum χ^2 values; i.e., the values obtained by using the parameter estimates listed in Table 3. The χ^2 values needed for significance at the 0.01 level are 26.22, 27.69, and 29.14 for 12, 13, and 14 degrees of freedom, respectively. To indicate the magnitude of discrepancy that produces a particular value of χ^2, Table 5 gives the observed and predicted response sequence probabilities for Experiment II. (We chose to display data for this experiment since it included the largest number of observations.)

Tables 4 and 5 demonstrate that certain models perform markedly better than do others. Neither the one-element nor the simple linear model yields accurate predictions. The sources of the disparity for these two models are about the same in all sets of data, and are indicated by Table 5. Especially prominent is the tendency for the linear model to predict too few sequences of all correct responses.

According to Table 4, the RTI model is consistently more accurate than the other two models which include linearity assumptions. Since the additional analyses to be reported corroborated this finding, we conclude that the simple linear and two-phase linear models (as well as the one-element model) are relatively inadequate. Hereafter, we shall restrict our attention chiefly to the remaining models.

Of the three-parameter models, the two-element model is less accurate than the LS-3 model in seven of the eight experiments. Both the LS-3 and LS-2 models do reasonably well. As Table 4 indicates, the number of data sets with significant χ^2's is less for these models than for any others. Also the values of χ^2 summed over data sets are lowest for the long-short models (see Table 4). The addition of the c parameter to the long-short formulation created only little improvement in the fits. This finding reflects the fact, mentioned earlier, that the estimate of c was usually close to 1.

To summarize Table 4, the long-short models are superior in predicting response sequence frequencies; of the remaining models, the RTI variant is most accurate. In the sequel, we shall be primarily concerned with testing these three models against other statistics.

TABLE 6

OBSERVED AND PREDICTED PROPORTIONS CORRECT ON TRIAL 2

Experiment	Observed	RTI	LS-3	LS-2
Ia	.679	.624	.665	.689
Ib	.597	.581	.586	.627
II	.601	.532	.598	.598
III	.531	.446	.519	.519
IV	.513	.487	.510	.540
Va	.446	.369	.435	.435
Vc	.544	.442	.538	.538
Ve	.660	.509	.618	.622

One factor that explains why the long-short models do better than the RTI model on the χ^2 measure becomes obvious when an inspection is made of the learning curves (Eq. 10) associated with the parameter estimates given in Table 3. All three models accurately predict the Pr (e_n) over trials except that the RTI model does rather poorly for trial 2. Table 6 presents the observed and predicted proportions correct on trial 2. For each of the eight sets of data, the discrepancy is greater for the RTI model than for the long-short models. The mean deviation between observed and predicted proportions is 0.07 for the RTI model against 0.01 for the long-short model.

Error Rate Conditionalized on Previous Errors

We now consider Pr $(e_{n+1} \mid e_n)$, the probability of an error on trial $n + 1$ conditional on an error on trial n. It will be seen that this statistic, although not independent of those discussed in the previous subsection, is quite useful in discriminating among various models. For example, in the one-element model

$$\Pr(e_{n+1} \mid e_n) = (1 - g)(1 - c),$$

which is constant over trials; whereas, for the simple linear model

$$\Pr(e_{n+1} \mid e_n) = (1 - g)(1 - \theta)^n$$

and decreases as n increases.

According to the RTI model, Pr $(e_{n+1} \mid e_n)$ must decrease as n increases. As indicated in Eq. 7, for the long-short formulation the trend of this conditional probability depends on the parameter values. When we plot the observed values of Pr $(e_{n+1} \mid e_n)$ as a function of n for each of our eight experiments the results are fairly decisive. For six of the eight curves, Pr $(e_{n+1} \mid e_n)$ clearly decreases as n increases. The exceptions are Experiments Va and Vc; in both of these cases the observed functions appear to be reasonably constant over trials. Also, Williams (1962) found that the probability of an error, conditionalized on no prior correct responses to that paired-associate item, decreased over trials. Using the parameter estimates given in Table 3, we find that the RTI model and the LS-3 model fit our observed Pr $(e_{n+1} \mid e_n)$ curves about equally well. If we compute the sum over trials of the absolute difference between predicted and observed values, then for Experiments III, IV, and Ve the RTI model yields a smaller sum than the LS-3 model, whereas the opposite is true for the other five experiments.

A strong prediction of the long-short formulation when $c = 1$ is that Pr $(e_{n+1} \mid e_n)$ is constant over trials; i.e.,

$$\Pr(e_{n+1} \mid e_n) = (1 - a)(1 - g)f.$$

Since this prediction was borne out in only two of the eight experiments, we are inclined to reject the LS-2 model as an adequate theory of paired-associate learning. However, in this regard it is interesting to note that the Vincent curve method suggested by Suppes and Ginsberg (1963) to test the stationarity prediction of the one-element model gives rise to ambiguous results when applied to the LS-2 model. For the LS-2 model we have stationarity after trial 1 and before the last error, but it is confounded by the probability of a correct response on trial 1. Specifically, if $\Pr(c_n \mid e'_m)$ is the probability of a correct response on trial n, given that the last error occurs on trial m ($m > n$), then for the LS-2 model

$$\Pr(c_n \mid e'_m) = \begin{cases} g, & \text{for} \quad n = 1 \\ 1 - f(1-g), & \text{for} \quad n > 1. \end{cases}$$

This result implies that a Vincent curve constructed by the methods prescribed by Suppes and Ginsberg will not be constant under the assumptions of the LS-2 model. Instead, it will exhibit an increase from the first part to later parts.

In the LS-3 model, the relation $c < 1$ changes the second equality in the above equation to an inequality; i.e.,

$$\Pr(c_n \mid e'_m) < 1 - f(1-g),$$

for $m > n$. This inequality follows from the fact that the subject can be in state U on trial n. After substituting $g = \frac{1}{3}$ for Experiments Ia, Ib, and II and $g = \frac{1}{4}$ for Experiments III, IV, and V and using the estimates of f from Table 3 we see that the predicted upper bounds on $\Pr(c_n \mid e'_m)$ are surprisingly low. The theoretical proportions in question range from 0.37 to 0.75 with a median of 0.46. To test whether the data satisfied the above inequality, we used the observed proportion correct on the trial immediately preceding the last precriterion error as an approximation to the observed maximum of $\Pr(c_n \mid e'_m)$. Except for Experiments III and Vc, the relevant observed proportion was quite near or below the predicted upper bound. In both of these data sets, the observed proportion exceeded that predicted by approximately 0.07. The import of this discrepancy is hard to ascertain. Considering the low predicted values, it is gratifying that the observed quantities did not further overshoot the predicted upper bounds. However, the error appears too large to be attributable to sampling fluctuations. Perhaps a decline in f after trial 5 contributes to the error in predictions for Experiments III and Vc. The next section traces the development of a model in which f does decrease over trials and some problems with this formulation are noted.

Trial-Dependent Forgetting Process

More generally, any Markov model with only one error state and constant transition and guessing probabilities predicts the stationarity of $\Pr(e_{n+1} \mid e_n)$. For example,

consider a model developed by Crothers (1963) that distinguishes between three states of learning; a guessing state ($\bar{C}$), a weak state of conditioning in which forgetting can occur (S), and a strong state of conditioning (L). The general formulation of his model is in terms of the following matrix and response probability vector:

$$\begin{array}{c} \\ L \\ S \\ \bar{C} \end{array} \begin{array}{c} \begin{array}{ccc} L & S & \bar{C} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ a & 1-a-b & b \\ c & 1-c-d & d \end{bmatrix} \end{array} \quad \begin{array}{c} \text{Pr (correct)} \\ \begin{bmatrix} 1 \\ 1 \\ g \end{bmatrix} \end{array} .$$

This model predicts a nonstationary Vincent curve, but (like the LS-2 model) it also predicts that Pr $(e_{n+1} \mid e_n)$ is constant over trials. In fact, our LS-2 model is a special case of the Crothers model. For when $c = 1$ it is no longer necessary to distinguish between states U and F in the matrix of Eq. 3, and therefore the process can be described as follows:

$$\begin{array}{c} \\ L \\ S \\ \bar{C} \end{array} \begin{array}{c} \begin{array}{ccc} L & S & \bar{C} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ a & (1-a)(1-f) & (1-a)f \\ a & (1-a)(1-f) & (1-a)f \end{bmatrix} \end{array} \quad \begin{array}{c} \text{Pr (correct)} \\ \begin{bmatrix} 1 \\ 1 \\ g \end{bmatrix} \end{array} . \tag{21}$$

Despite an unrealistic prediction (i.e., that Pr $(e_{n+1} \mid e_n)$ is constant over trials), the LS-2 model describes many aspects of the eight data sets remarkably well, as indicated by the results in Tables 4 and 5. Hence, for the moment it seems worthwhile to retain the basic structure of the LS-2 model and determine what can be gained by pursuing a generalization of the forgetting process that would permit Pr $(e_{n+1} \mid e_n)$ to decrease over trials. We now examine one such generalization as an alternative to the LS-3 and RTI models.

Under the assumptions of the LS-2 model, if item i is reinforced it passes into state L with probability a or into state S with probability $1 - a$. Once in L it is trapped there; but if in S it may move back to $\bar{C}$. That is, other stimuli intervene from one presentation of item i to its next presentation and during this period there is probability f that forgetting will take place (i.e., item i will pass from state S to $\bar{C}$). Thus the forgetting process depends only on the number of intervening stimuli and is independent of the stage of learning. One obvious generalization is to assume that the likelihood of forgetting is not simply a function of the number of intervening items, but depends on the number of intervening items that have not already been learned. With this modification the transition probabilities become functions of the trial number, and the matrix in Eq. 21 is rewritten as

$$\begin{array}{c} \\ L \\ S \\ \bar{C} \end{array} \begin{array}{c} \begin{array}{ccc} L & S & \bar{C} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ a & (1-a)(1-F_n) & (1-a)F_n \\ a & (1-a)(1-F_n) & (1-a)F_n \end{bmatrix} \end{array} , \tag{22}$$

where F_n is a function of the number of unlearned items that intervene from the nth presentation of item i to its $n + 1$st presentation. Let us assume that each unlearned stimulus gives rise to complete forgetting of item i with probability f'. Thus, if there are k unlearned stimuli presented between the nth and $n + 1$st presentation of item i, then

$$F_n = 1 - (1 - f')^k.$$

By an unlearned stimulus item, we mean an item not already in state L. Further, for a list length of $X + 1$ items the expected number of unlearned items that intervene between the nth and $n + 1$st presentation of a particular item is simply

$$X(1 - a)^n.$$

Using this expected value as an *approximation* to the actual number of unlearned items that intervene from the nth to the $n + 1$st presentation of a given stimulus item we can write

$$F_n = 1 - (1 - f')^{X(1-a)^n}.$$

Finally, to attain more generality, let us assume that forgetting also can occur during the intertrial interval with probability f. Including this factor in the forgetting process yields the following expression:

$$F_n = 1 - (1 - f)(1 - f')^{X(1-a)^n}. \tag{23}$$

The model described by Eqs. 22 and 23 has three parameters: f, f', and a. Also, the model takes explicit account of the list length variable and the intertrial interval.

Henceforth, this model will be referred to as the trial-dependent-forgetting process (TDF model). Of course, for the TDF model, $\Pr(e_{n+1} \mid e_n)$ is a decreasing function of the trial number; i.e.,

$$\Pr(e_{n+1} \mid e_n) = (1 - g)(1 - a)F_n .$$

When $f' = 0$, the model reduces to the LS-2 process.

Using Eqs. 22 and 23, we generated expressions (comparable to those given by Eq. 19) for four-tuple response sequences. Minimum χ^2's were then computed for the data reported in Table 2; in carrying out the minimizations account was taken of the list length variable $X + 1$ as given in Table 1. Two sets of minimizations were run: one involved estimating the three parameters a, f, and f'; the other involved estimating only a and f' (under the assumption that $f = 0$). The resulting χ^2 values and associated parameter estimates are given in Table 7. The two-parameter case yields a total χ^2 value of 205.92, which compares favorably with the total χ^2's for the other two-parameter models (see Table 4). The three parameter case yields a total χ^2 of 137.55 which is virtually identical to the total χ^2 value for the LS-3 model. Thus, in terms of

TABLE 7

PARAMETER ESTIMATES AND MINIMUM χ^2 VALUES FOR THE TDF MODEL

Experiment	Three parameter case				Two parameter case		
	χ^2	a	f'	f	χ^2	a	f'
Ia	5.18[a]	.328	.094	.391	6.44[a]	.289	.141
Ib	15.07[a]	.281	.086	.266	15.36[a]	.266	.102
II	3.71[a]	.242	.016	.766	15.55[a]	.219	.156
III	33.02	.188	0	.789	44.40	.164	.141
IV	8.92[a]	.250	.094	.398	9.80[a]	.242	.133
Va	24.41[a]	.109	.016	.836	28.92	.102	.414
Vc	27.12	.156	0	.727	39.24	.141	.320
Ve	20.12[a]	.258	0	.680	46.21	.211	.281
Total χ^2	137.55				205.92		

[a] Not significant at .01 level.

these analyses it is difficult to choose between the LS-3 model (which postulates a constant forgetting process and a coding operation), and the three parameter version of the TDF model (which does not postulate a coding operation but makes the forgetting process time dependent). However, the fact that the TDF model is non-Markovian greatly enhances the difficulty of performing derivations (e.g. of distributions and expectations) for an infinite sequence of trials. Also, other ways of introducing processes with trial-dependent parameters can be suggested that are *a priori* about as plausible as the approach outlined. These reasons lead us to doubt that the TDF model is among the more promising conceptualizations.

TRIAL NUMBER OF THE LAST ERROR AND TOTAL ERRORS

Our final test of the LS-3 model consisted in predicting the expected total errors per subject-item, and the distribution and expectation of the trial number of the last error. The theoretical values for Experiments Ia and Ib were obtained by substituting the parameter estimates given in Table 3 into Eqs. 11 and 14; the predicted expected trial of last error was approximated by direct computation based on the first eleven terms of the theoretical probability distribution, which summed to approximately 0.99.

As one would expect, there is good agreement between the observed and predicted mean trial number of last error. The one serious discrepancy disclosed by Table 8 is that the first two terms of the probability distribution of k (the trial number of last error) are inadequately predicted in Experiment Ia. According to the model, the distribution should be peaked at $k = 1$, whereas the observed proportions attain a

TABLE 8

OBSERVED AND PREDICTED (LS-3 MODEL) VALUES FOR EXPERIMENT Ia AND Ib

	Exp. Ia		Exp. Ib	
	Obs.	Pred.	Obs.	Pred.
Expected total errors	1.52	1.54	1.65	1.79
Expected trial of last error	1.76	2.05	2.08	2.45
Probability of last error on trial k				
$k = 0$	.27	.17	.16	.14
$k = 1$	.24	.34	.26	.27
$k = 2$	.19	.19	.19	.19
$k = 3$	.13	.11	.17	.13
$k = 4$	.10	.07	.11	.09
$k = 5$	.04	.05	.07	.06
$k = 6$	.02	.03	.02	.04
$k = 7$	.01	.02	.00	.03

maximum at $k = 0$ and decrease monotonically over k. There is also a slight tendency to underestimate the peaking at $k = 1$ in Experiment Ib.

Further work is required to determine the source of the discrepancy in the initial terms of this distribution. At present, it is uncertain to what extent the deviation indicates an actual departure from the assumed learning mechanism. An alternative explanation is that the learning parameters increase (or the forgetting parameter decreases) over trials. If this were the case, the estimates based on trials 2-5 would predict slower learning than that observed.

VI. DISCUSSION AND CONCLUSIONS

The results of our analyses indicate that five of the seven models tested yield relatively unsatisfactory predictions for paired-associate learning under the experimental conditions described. One immediate question is why the one-element model was consistently inaccurate. At first glance, the reply might be that we estimated parameters and tested predictions in a different fashion than did Bower (1961). However, the model fails in Experiment Ia where well over 95% of the errors are included in the four-tuple analysis; further, for our experiments the $\Pr(e_{n+1} \mid e_n)$ curves do not exhibit the stationarity predicted by the one-element model. Therefore, it seems more likely that differences in experimental method are responsible for the inadequate performance of this model. The most important procedural difference

appears to be that the number of response alternatives was two in Bower's study and three or four in the experiments reported here.

Of the models considered in this paper the LS-3 and the RTI models seem to warrant first consideration for future experimental tests and theoretical development. The findings in favor of the former model are not conclusive, but its parameters have been identified more closely with psychological processes. Such interpretations are helpful in suggesting how the parameter estimates should change under various experimental manipulations. Experiments Ia and Ib provide evidence on this point for the list length variable. As another example, it would be of interest to determine if the forgetting parameter f is invariant when the type of paired-associate stimulus is changed from one experimental condition to another. Also, perhaps rehearsal of irrelevant material during the intertrial interval would affect only the forgetting parameter.

Another direction for further investigation involves improving the minimum χ^2 technique of parameter estimation, especially when the data in question display a high proportion of errors after trial 5. We can write equations for seven-tuples (trials 2-8) without much difficulty. The derivation depends upon finding the state probabilities of trial 5 conditional on a particular sequence on trials 2-4, and then using Eq. 19. Beyond seven-tuples, however, the derivations become quite cumbersome. Further, it is pertinent to know how well the parameter estimates based on four-tuple data will predict statistics that include data from other trials. Our preliminary work on this point involved using estimates generated from the χ^2 minimum method to predict the distribution of the trial number of the last error and total errors. Obviously tests of this type need to be extended to other statistics and data sets.[6]

Further work also should include examining the extent to which the LS-3 model can be altered without reducing its goodness-of-fit. The remaining remarks are aimed at suggesting what can be done in explorations of this nature.

In our original formulation of the long-short model it seemed natural to view forgetting as an event that influenced response probability by changing the learning state. On the other hand, forgetting can be interpreted as affecting the response probability directly, without producing a state transition. That is, the LS-3 model can be rewritten by collapsing states S and F and making the response probability in the single intermediate state (let us call this state SF) a function of the forgetting parameter. For the transition matrix and response probability vector we have

$$\begin{array}{c} \\ L \\ SF \\ U \end{array} \begin{array}{c} \begin{array}{ccc} L & SF & U \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ a & 1-a & 0 \\ ca & c(1-a) & 1-c \end{bmatrix} \end{array} \begin{array}{c} \text{Pr (correct)} \\ \begin{bmatrix} 1 \\ 1-f+fg \\ g \end{bmatrix} \end{array}. \tag{24}$$

[6] For a discussion of this general topic the reader is referred to an article by Sternberg (1963, pp. 89-99).

This representation of the LS-3 model is algebraically identical to the original formulation given in Eq. 3. Hence both formalizations yield identical predictions for response events and any preference for one over the other would seem to derive from their respective heuristic merits. For example, one way of treating response latency within the framework of the LS-3 model is to postulate a latency distribution associated with each of the learning states. For the data we have seen, there is reason to believe that it would be necessary to postulate four such distributions to give an accurate account of latency measures. Hence, if one were to take this approach to the analysis of latency data, then the formalization given by Eq. 3 would be more natural than that of Eq. 24.

The three-state representation given in Eq. 24 suggests two ways of modifying the LS-3 model that leave the forgetting mechanism unchanged but affect other aspects of the process. One such modification is to assume that passage into state L occurs with probability a on any trial regardless of the current state. In this variation of the LS-3 model, the transition matrix and response probability vector (for the representation in which states S and F are collapsed) are as follows:

$$\begin{array}{c} \\ L \\ SF \\ U \end{array}\begin{array}{c} \begin{array}{ccc} L & SF & U \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ a & 1-a & 0 \\ a & c-a & 1-c \end{bmatrix} \end{array} \quad \begin{array}{c} \text{Pr (correct)} \\ \begin{bmatrix} 1 \\ 1-f+fg \\ g \end{bmatrix} \end{array}, \tag{25}$$

where $c \geqslant a$. Applying this model to the four-tuple response data yields the parameter estimates and associated χ^2 values shown in Table 9. In terms of the total χ^2 measure, this model is as accurate as the original LS-3 model.

TABLE 9

PARAMETER ESTIMATES AND MINIMUM χ^2 VALUES FOR THE MODEL DESCRIBED BY EQ. 25

Experiment	Parameter			χ^2
	a	f	c	
Ia	.336	.677	.761	6.15
Ib	.227	.365	.475	13.50
II	.250	.898	.994	3.78
III	.188	.781	.950	32.92
IV	.234	.541	.455	8.34
Va	.109	.844	.993	24.43
Vc	.156	.729	.980	27.06
Ve	.266	.688	.995	20.36
		Total χ^2		136.54

Another variation of the LS-3 model suggested by an inspection of Eq. 24 is to assume that in state U there is probability c of moving to state L, whereas in either state S or F there is probability a of moving to state L. More specifically, we have in mind a three parameter variation on the LS-3 model with the following transition matrix and response vector:

$$\begin{array}{c} \\ L \\ SF \\ U \end{array}\begin{array}{c} \begin{array}{ccc} L & SF & U \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ a & 1-a & 0 \\ c & (1-c)\,a & (1-c)(1-a) \end{bmatrix} \end{array}\quad \begin{array}{c} \text{Pr (correct)} \\ \begin{bmatrix} 1 \\ 1-f+fg \\ g \end{bmatrix} \end{array}. \tag{26}$$

Applying this model to the four-tuple data yields the parameter estimates and χ^2 values given in Table 10. For this model the total χ^2 is almost twice that obtained for the original LS-3 model.

TABLE 10

PARAMETER ESTIMATES AND MINIMUM χ^2 VALUES FOR THE MODEL DESCRIBED BY EQ. 26

Experiment	Parameter			χ^2
	a	f	c	
Ia	.391	.445	.297	8.05
Ib	.398	.375	.156	12.49
II	.242	.563	.234	17.47
III	.227	.427	.156	46.20
IV	.250	.469	.227	8.68
Va	.188	.479	.063	37.30
Vc	.273	.344	.078	57.02
Ve	.289	.511	.258	65.61
	Total χ^2			252.82

Our aim in citing these two variations on the original LS-3 model has been to indicate the degree of discrimination that is attainable among related models by using four-tuple data. The version given in Eq. 25 fits as well as the original LS-3 model. This fact is grounds for affirming the caution voiced in connection with the TDF model; namely, that a single accurate model and its interpretation should not be prematurely accepted. On the other hand, the model given in Eq. 26 was far less adequate, so we can sharpen an earlier conclusion. The relatively inaccurate fit of the two-element model (Section V) led us to assert that not all plausible three-parameter models yield equally accurate fits; now we note that considerable discrimination can be achieved within a special family of three-parameter models.

APPENDIX

The aim of this section is to derive necessary and sufficient conditions under which the members of a restricted class of Markov models yield algebraically equivalent expressions for probabilities of response n-tuples. Of course, they will then yield equivalent equations for summary statistics (expected total errors, etc.). The results that we present in this section are in the same vein as those obtained by Greeno (1964).

We consider two models, each with states L (long-term), S (short-term), and U (unconditioned). Throughout this section we also assume that the subject begins in state U and that the probability of a correct response equals g in state U and equals one in both states S and L. Let the state transitions be described by the matrices $[A]$ and $[B]$ for the first and second models, respectively, where

$$[A] = \begin{bmatrix} 1 & 0 & 0 \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, \tag{A.1}$$

$$[B] = \begin{bmatrix} 1 & 0 & 0 \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix}.$$

What conditions must the b_{ij} entries satisfy in order that Models A and B are isomorphic in their predictions of response sequence probabilities? A trivial sufficient condition is that $a_{ij} = b_{ij}$ for each i and j. What is more interesting is that Models A and B are equivalent if and only if the following equalities hold:

$$a_{33} = b_{33} \tag{A.2}$$

$$a_{22} = b_{22} \tag{A.3}$$

$$a_{32} \cdot a_{23} = b_{32} \cdot b_{23}. \tag{A.4}$$

Requirement (A.2) follows from the fact that $\Pr(e_2) = (1 - g)\, a_{33}$ in Model A and $\Pr(e_2) = (1 - g)\, b_{33}$ in Model B. Likewise, the equations for $\Pr(e_3)$ can be used to show that Eq. A.4 must hold in order that the models be isomorphic. Then by substitution of Eqs. A.2 and A.4 into the two equations for $\Pr(e_4)$, we derive the relation $a_{22} = b_{22}$ as another necessary condition for equivalence of the models.

To prove that Eqs. A.2-A.4 are sufficient to ensure equivalence, it is only necessary to show that any n-tuple of responses can be written in the form

$$\Pr(n\text{-tuple}) = f(g, n, a_{22}, a_{33}, a_{23} \cdot a_{32}), \tag{A.5}$$

i.e., that a_{23} and a_{32} enter only as the product $a_{23}a_{32}$. We assume that the subject was in U on the trial before the start of the n-tuple. Of course, the number of the trial in U is immaterial.

It suffices to prove Eq. A.5 for each of the $(2^n - 1)$ n-tuples which include at least one error, since then this equation must hold for the one remaining n-tuple. For expositional convenience, we indicate the method of proof by taking a particular n-tuple, say $c_2c_3e_4c_5c_6$. Here $n = 5$ and

$$\Pr(c_2c_3e_4c_5c_6) = \Pr(c_6c_5 \mid e_4) \Pr(c_2c_3e_4). \tag{A.6}$$

Now we argue by induction on n. It is obvious that Eq. A.5 holds for $n = 1$; the possible 1-tuples are $\Pr(c_2) = \Pr(c_2 \mid e_1)$ and $\Pr(e_2) = \Pr(e_2 \mid e_1)$. The second principle of finite induction allows us to assume (A.5) for $k < n$. In particular,

$$\Pr(c_6c_5 \mid e_4) = \Pr(c_3c_2 \mid e_1)$$

and we assume that $\Pr(c_3c_2 \mid e_1)$ can be written in the form of (A.5). The next step is to show that (A.5) also holds for the second factor on the right side of (A.6). A sequence ending in an error can occur in one of two ways. Either the subject remains in U throughout the sequence, or he passes to S and eventually returns to U. In order to end the sequence in U, each transition from U to S must be followed by a transition from S to U, which in turn implies that (A.5) holds. Therefore the second factor on the right side of (A.6) satisfies (A.5). Having established that each of the two factors on the right side of (A.6) fulfills (A.5), it follows that the product of the two factors can be written in the form of (A.5).

In general, all but one of the n-tuples starting on the trial after a response in U can be decomposed into an $(n - k)$-tuple which ends in an error and a k-tuple which begins on the trial after an error. The foregoing reasoning shows that the probability of each of the tuples satisfies (A.5). Hence the product of the two probabilities satisfies (A.5). Q.E.D.

A pair of transition matrices that indeed satisfies Eqs. A.2-A.4 is

$$[T_1] = \begin{bmatrix} 1 & 0 & 0 \\ 0 & a & 1-a \\ c & d & 1-c-d \end{bmatrix}, \tag{A.7}$$

$$[T_2] = \begin{bmatrix} 1 & 0 & 0 \\ x & y & 1-x-y \\ 0 & z & 1-z \end{bmatrix}. \tag{A.8}$$

After making the substitutions

$$a = y, \qquad 1 - c - d = 1 - z, \qquad \text{and} \qquad (1 - a)\, d = (1 - x - y)\, z,$$

we have

$$[T_2] = \begin{bmatrix} 1 & 0 & 0 \\ \dfrac{(1-a)\,c}{c+d} & a & \dfrac{(1-a)\,d}{c+d} \\ 0 & c+d & 1-c-d \end{bmatrix}.$$

Hence Model T_2, which assumes that direct transitions do not occur from U to L, is indistinguishable at the response level from Model T_1, which assumes that immediate transitions from S to L are impossible. Estimates of the corresponding off-diagonal entries in T_1 and T_2 will differ numerically, but the predicted n-tuple probabilities will be identical.

The results summarized in Eqs. A.2-A.4 are helpful in understanding when the addition of parameters to a 3-state model will lower the goodness-of-fit χ^2. For example, let us compare the 3-parameter Model T_1 (cf. Eq. A.7) with the LS-2 model. The transition matrix for the latter has the form

$$[\text{LS-2}] = \begin{bmatrix} 1 & 0 & 0 \\ a' & b' & 1-a'-b' \\ a' & b' & 1-a'-b' \end{bmatrix}.$$

Now denote by $\hat{a}$, $\hat{c}$, and $\hat{d}$ the minimum χ^2 estimates of a, c, and d in Eq. A.7. What algebraic relationship must hold among the $\hat{a}$, $\hat{c}$, and $\hat{d}$ values in order that the LS-2 model yield as good a fit as the T_1 (or T_2) model?

By Eqs. A.2 and A.3 we choose a' and b' so as to satisfy the equations $b' = a$ and $1 - a' - b' = 1 - c - d$. Referring to Eq. A.7, we see that Eq. A.4 holds if and only if $\hat{a}(1 - \hat{c} - \hat{d}) = (1 - \hat{a})\,\hat{d}$. Therefore we begin by using the computer search routine to find $\hat{a}$, $\hat{c}$, and $\hat{d}$. If the two sides of the foregoing equation are approximately equal, then the 2-parameter LS-2 model will fit virtually as well as the 3-parameter model.

In practice, the goodness-of-fit χ^2 is diminished only slightly when we go from the LS-2 model to the T_1 (or T_2) model. Table 11 compares these χ^2 values for the experiments described in the body of this report. Except for Experiment Ve, the reduction in χ^2 is minimal.

No further improvement in the goodness-of-fit χ^2 is possible by going from a 3-parameter model (T_1 or T_2) to a 4-parameter model. Perhaps the truth of this statement is obvious from Eqs. A.2-A.4, but we shall demonstrate it explicitly. Let $\hat{a}_{ij}$ represent the minimum χ^2 estimate of a_{ij} for the 4-parameter model in (A.1). Having the $\hat{a}_{ij}$ values at hand, we can produce the same minimum χ^2 value from the 3-para-

TABLE 11

MINIMUM χ^2 VALUES FOR THE TWO AND THREE PARAMETER LS-MODEL

Experiment	LS-2	T_1 and T_2
Ia	6.75	6.61
Ib	19.69	18.37
II	3.73	2.80
III	33.02	31.78
IV	12.32	10.79
Va	24.41	23.98
Vc	27.12	24.10
Ve	20.12	14.56
Total χ^2	147.16	132.99

meter Model T_1. Equations A.2 and A.3 tell us to choose $\hat{a}$, $\hat{c}$, and $\hat{d}$ to meet the conditions

$$\hat{a} = \hat{a}_{22},$$

and

$$1 - \hat{c} - \hat{d} = \hat{a}_{33}.$$

Then by (A.4) and (A.7) we take

$$\hat{d} = \frac{\hat{a}_{23}\hat{a}_{32}}{1 - \hat{a}} = \frac{\hat{a}_{23}\hat{a}_{32}}{1 - \hat{a}_{22}}.$$

Because we can always pick $\hat{a}$, $\hat{c}$, and $\hat{d}$ so that these three conditions are fulfilled, the 3-parameter Model T_1 (or T_2) is isomorphic to the 4-parameter model in Eq. A.1. For expository purposes it was convenient to regard the $\hat{a}_{ij}$ as known prior to the calculation of $\hat{a}$, $\hat{c}$, and $\hat{d}$; however, by the above argument, there would be no point in finding the $\hat{a}_{ij}$ estimates.

ACKNOWLEDGMENTS

The authors wish to express their indebtedness to W. K. Estes, J. L. Myers, and P. Suppes for their valuable criticism and encouragement in the early stages of this research.

REFERENCES

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. Pp. 121-268.

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

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

BUSH, R. R., AND MOSTELLER, F. *Stochastic models for learning*. New York: Wiley, 1955.

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.

CRAMÉR, H. *Mathematical methods of statistics*. Princeton: Princeton Univ. Press, 1951.

CROTHERS, E. J. Paired-associate learning with compound response. *J. verb. Learn. verb. Behav.*, 1962, **1**, 66-70.

CROTHERS, E. J. Markov models for learning with inter-trial forgetting. *Tech. Rep. No. 53*, Institute for Mathematical Studies in the Social Sciences, Stanford Univ., 1963.

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

ESTES, W. K. New developments in statistical behavior theory: differential tests of axioms for associative learning. *Psychometrika*, 1961, **26**, 73-84.

ESTES, W. K., HOPKINS, B. L., AND CROTHERS, E. J. All-or-none and conservation effects in the learning and retention of paired associates. *J. exp. Psychol.*, 1960, **60**, 329-339.

GREENO, J. G. Markovian learning processes with identifiable states: I. General considerations and application to all-or-none learning. Mimeographed paper, Indiana Univ., 1964.

HANSEN, D. N. Paired-associate learning with young children. Unpublished doctoral dissertation, Stanford Univ., 1963.

HELLYER, S. Supplementary report: Frequency of stimulus presentation and short-term decrement in recall. *J. exp. Psychol.*, 1962, **64**, 650.

LAWRENCE, D. H. The nature of a stimulus: some relationships between learning and perception. In S. Koch (Ed.) *Psychology: a study of a science*, Vol. 5. New York: McGraw-Hill, 1963. Pp. 179-212.

MELTON, A. W. Comments on Professor Peterson's paper. In Cofer, C. N., and Musgrave, Barbara (Eds.) *Verbal behavior and learning*. New York: McGraw-Hill, 1963a. P. 355.

MELTON, A. W. Implications of short-term memory for a general theory of memory. *J. verb. Learn. verb. Behav.*, 1963b, **2**, 1-21.

MURDOCK, B. B. The retention of individual items. *J. exp. Psychol.*, 1961, **62**, 618-625.

NORMAN, M. F. A two-phase model and an application to verbal discrimination learning. In R. C. Atkinson (Ed.) *Studies in mathematical psychology*. Stanford: Stanford Univ. Press, 1964. Pp. 173-187.

NORMAN, M. F. Incremental learning on random trials. *J. math. Psychol.*, 1964, **1**, 336-350.

PETERSON, L. R., AND PETERSON, MARGARET J. Short-term retention of individual verbal items. *J. exp. Psychol.*, 1959, **58**, 193-198.

PETERSON, L. R., SALTZMAN, DOROTHY, HILLNER, K., AND LAND, VERA. Recency and frequency in paired-associate learning. *J. exp. Psychol.*, 1962, **63**, 396-403.

STERNBERG, S. Stochastic learning theory. In R. D. Luce, R. R. Bush, and E. Galanter (Eds.) *Handbook of mathematical psychology*, Vol. 2. New York: Wiley, 1963. Pp. 1-120.

SUPPES, P., AND GINSBERG, ROSE. A fundamental property of all-or-none models. *Psychol. Rev.*, 1963, **70**, 139-161.

WILLIAMS, JOANNA P. A test of the all-or-none hypothesis for verbal learning. *J. exp. Psychol.*, 1962, **64**, 158-166.

RECEIVED: November 11, 1963