INCREMENTAL LEARNING ON RANDOM TRIALS

by

M. Frank Norman[1]

TECHNICAL REPORT NO. 62

December 9, 1963

PSYCHOLOGY SERIES

INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES

STANFORD UNIVERSITY

Stanford, California

Abstract

When a subject is required to learn to make a certain response in the course of a sequence of consistently corrected trials, and when it is not clear which of success or failure is more efficacious for learning, the possibility suggests itself that acquisition might sometimes proceed incrementally but with a certain probability, constant over trials, that no learning occurs on a trial. In this paper many of the properties which one would expect acquisition of this type to possess if the incremental process were linear in the response probabilities are derived, and details of an application of this model to a paired-associate experiment are presented.

In a previous paper (Norman, 1963) an attempt was made to interpolate between the all-or-none model (see Bower, 1961) and the single-operator linear model (see Bush and Sternberg, 1959) by means of a "two-phase" model. In this model a random number of trials on which no change in response probability occurred preceded the beginning of the linear incremental learning process described by the single-operator model. In the present paper an interpolation of a slightly different kind is discussed; all applications of a linear operator to response probabilities are preceded by a random number of trials on which no learning occurs. This model will be referred to below as the random-trial incremental model. Many of its mathematical properties will be derived, and data from a paired-associate learning experiment will be presented to illustrate its usefulness.

A delightful feature of the single-operator and all-or-none models is their mathematical simplicity. It will emerge from the development below that this simplicity is shared in significant measure by the random-trial incremental model.

The Model

Let us assume that a subject's probability $p_{\vec{w}n}$ of making the A_2 (error) response on trial n satisfies the stochastic difference equation

$$p_{\vec{w}n+1} = a^{y_{\vec{w}n}} p_{\vec{w}n} \qquad n = 1,2,\ldots, \tag{1}$$

and the initial condition

$$p_{\underline{w}1} = \underline{p}$$

where $\{y_{\underline{w}n}\}$ is a sequence of identically and independently distributed random variables with

(2) $$\underline{P}(y_{\underline{w}1} = 1) = \underline{c}$$

and

$$\underline{P}(y_{\underline{w}1} = 0) = 1 - \underline{c} \ .$$

It follows immediately that, for $\underline{n} \geq 1$,

(3) $$p_{\underline{w}n} = \underline{a}^{Y_{\underline{w}n-1}} \underline{p}$$

where

$$Y_{\underline{w}k} = \sum_{\underline{i}=1}^{k} y_{\underline{w}i} \qquad \text{for } \underline{k} \geq 1$$

and

$$Y_{\underline{w}0} = 0 \ .$$

For $\underline{k} \geq 1$, $Y_{\underline{w}k}$ clearly has the binomial distribution with parameters $\underline{k}$ and $\underline{c}$.

We must, of course, have $0 \leq \underline{a}, \underline{c}, \underline{p} \leq 1$. The reader will note that, when $\underline{a} = 0$, the random-trial incremental model reduces to the all-or-none model, whereas, if $\underline{c} = 1$, it reduces to the single-operator linear model. If the sequence $\{y_{\underline{w}n}\}$ is interpreted as the sequence of reinforcement indicator random variables for a non-contingent reinforcement schedule such that the probability of reinforcement

of A_1 is c on every trial then the random-trial incremental model is just the identity-operator model (see Bush and Mosteller, 1955) used with such a schedule. Some of the mathematical results obtained below are new in this context and it is hoped that they will be useful in this connection even though this was not the motivation for the research discussed in this paper. While the mathematical work was being done I was thinking of the sequence $\{y_{wn}\}$ as unobservable, and I was not able to identify the random process embodied in this sequence with any general psychological process or observable aspect of the experimental situation in the application of the model to be presented below.

The Basic Theorem

Most of the properties of the random-trial incremental model which I will discuss in this paper are readily obtained from the following fundamental result:

Let k be a positive integer and let $i_1, i_2, \ldots, i_k$ be a strictly increasing finite sequence of positive integers. Then

$$(4) \qquad P(x_{wi_1} = 1, x_{wi_2} = 1, \ldots, x_{wi_k} = 1) = p^k \prod_{j=1}^{k} b_j^{i_{k-j+1} - i_{k-j}}$$

where

$$b_j = a^j c + (1-c) \qquad \text{for } j = 1,2,\ldots,$$

$i_0 = 1$, and x_{wm} is the error indicator random variable for trial m.

The following proof proceeds by induction on k.

1° If $k = 1$, (4) reduces to

$$P(x_i = 1) = p\, b_1^{i-1} \tag{5}$$

That this is correct is seen as follows:

$$P(x_i = 1) = \sum_{n=0}^{i-1} P(x_i = 1 \mid Y_{i-1} = n) P(Y_{i-1} = n)$$

$$= \sum_{n=0}^{i-1} a^n p \binom{i-1}{n} c^n (1-c)^{i-1-n}$$

$$= p(ac + (1-c))^{i-1} = p\, b_1^{i-1}$$

by the binomial theorem.

2° Suppose that (4) is correct for any collection of k indices $1 \le m_1 < m_2 < \ldots < m_k$ and all parameter values $0 \le a, c, p \le 1$. (In what follows, I will indicate the dependence of P on the initial error probability explicitly by means of a subscript.) Let some set $1 \le i_1 < i_2 < \ldots < i_{k+1}$ of $k+1$ indices and some parame values a, c, and p be given. Then

$$P_p(x_{i_1} = 1, \ldots, x_{i_k} = 1, x_{i_{k+1}} = 1)$$

$$= \sum_{m=0}^{i_1-1} P_p(x_{i_1} = 1, \ldots, x_{i_{k+1}} = 1 \mid Y_{i_1-1} = m) P(Y_{i_1-1} = m)$$

$$= \sum_{m=0}^{i_1-1} P_{a^m p}(x_1 = 1, x_{i_2-i_1+1} = 1, \ldots, x_{i_{k+1}-i_1+1} = 1) P(Y_{i_1-1} = m)$$

for it is a property of the random-trial incremental model that after the error probability on trial i_1 has been specified the sequence $x_{i_1}, x_{i_1+1}, x_{i_2+2}$ is stochastically identical to the sequence x_1, x_2, x_3 with the specified error probability on trial 1. Noting next that x_1 is independent of the x_i's with $i > 1$ we obtain

$$P_p(x_{i_1} = 1, \ldots, x_{i_k} = 1, x_{i_{k+1}} = 1) =$$

$$= \sum_{m=0}^{i_1-1} P_{a^m p}(x_1 = 1) P_{a^m p}(x_{i_2-i_1+1} = 1, \ldots, x_{i_{k+1}-i_1+1} = 1) P(Y_{i_1-1} = m) =$$

$$= \sum_{m=0}^{i_1-1} a^m p (a^m p)^k \left(\prod_{j=1}^{k} b_j^{\ell_{k-j+1}-\ell_{k-j}} \right) \binom{i_1-1}{m} c^m (1-c)^{i_1-1-m}$$

by the induction hypothesis where $\ell_h = i_{h+1} - i_1 + 1$

$$= p^{k+1} \left(\prod_{j=1}^{k} b_j^{i_{k+1-j+1}-i_{k+1-j}} \right) \sum_{m=0}^{i_1-1} \binom{i_1-1}{m} (a^{k+1} c)^m (1-c)^{i_1-1-m}$$

$$= p^{k+1} \prod_{j=1}^{k+1} b_j^{i_{k+1-j+1}-i_{k+1-j}} .$$

An application of the principle of induction completes the proof of (4).

Corollaries of the Basic Theorem

Let J_w be the number of errors before the first success. Then J_w is nonnegative and for $k \geq 1$, $P(J_w \geq k) = P(x_{w1} = 1, \ldots, x_{wk} = 1)$. Therefore,

(6) $$P(J \geq k) = p^k \prod_{j=1}^{k-1} b_j .$$

For $k \geq 1$ let $u_k = \sum_{n=1}^{\infty} \prod_{i=0}^{k-1} x_{n+i}$ be the total number of k tuples of consecutive errors in an infinite sequence of trials. Then

$$E(u_k) = \sum_{n=1}^{\infty} P(x_n = 1, \ldots, x_{n+k-1} = 1)$$

$$= \sum_{n=1}^{\infty} p^k \left(\prod_{j=1}^{k-1} b_j \right) b_k^{n-1} .$$

Thus

(7) $$E(u_k) = \frac{p^k \prod_{j=1}^{k-1} b_j}{1 - b_k}$$

(where $\prod_{j=1}^{0} b_j$ is defined to be 1; u_1, the total number of error in an infinite sequence of trials will also be denoted by T in what follows).

For $k \geq 1$, let $c_k = \sum_{n=1}^{\infty} x_n x_{n+k}$ be the autocovariance of errors of lag k (clearly $c_1 = u_2$). Then

$$E(c_k) = \sum_{n=1}^{\infty} P(x_n = 1,\ x_{n+k} = 1)$$

$$= \sum_{n=1}^{\infty} p^2\, b_1^k\, b_2^{n-1} \quad .$$

Therefore

$$E(c_k) = \frac{p^2\, b_1^k}{1 - b_2} \quad . \tag{8}$$

Clearly

$$T^2 = \sum_{i=1}^{\infty} x_i^2 + 2 \sum_{k=1}^{\infty} \sum_{n=1}^{\infty} x_n\, x_{n+k} = T + 2 \sum_{k=1}^{\infty} c_k \quad .$$

Thus

$$E(T^2) = E(T) + 2 \sum_{k=1}^{\infty} E(c_k)$$

$$= E(T) + 2 \sum_{k=1}^{\infty} \frac{p^2\, b_1^k}{1 - b_2}$$

by (8)

$$= E(T) + \frac{2p^2}{1 - b_2} \frac{b_1}{1 - b_1} = E(T) + \left(\frac{p}{1 - b_1}\right) \left(\frac{2}{p}\right) \left(\frac{p^2\, b_1}{1 - b_2}\right) \quad .$$

We conclude that

$$E(T^2) = E(T) + \frac{2}{p} E(T) E(u_2) \quad . \tag{9}$$

Similar computations yield the formulas

$$E(T\, u_2) = E(T) E(u_2) \left[\frac{2}{p} (1 - b_1) + \frac{b_2(1 - b_2)}{1 - b_3} + \frac{b_2(1 - b_1)}{1 - b_3} \right] \tag{10}$$

and

$$(11) \qquad E(u_2^2) = E(u_2) + 2E(u_3) + \frac{2}{1-b_2} E(u_4)$$

which we will use in evaluating our estimation procedure below.

For $n \geq 1$ denote by $T^{(n)}$ the total number of errors after the n th trial in an infinite sequence of trials, i.e.,

$T^{(n)} = \sum_{k=n+1}^{\infty} x_k$. Then

$$E(T^{(n)} | x_n = 1) = \sum_{k=n+1}^{\infty} \frac{P(x_k = 1, x_n = 1)}{P(x_n = 1)}$$

$$= \sum_{k=n+1}^{\infty} \frac{p^2 b_1^{k-n} b_2^{n-1}}{p b_1^{n-1}}$$

$$= p \left(\frac{b_2}{b_1}\right)^{n-1} \frac{b_1}{1-b_1} .$$

Thus

$$(12) \qquad E(T^{(n)} | x_n = 1) = E(T) b_1 \left(\frac{b_2}{b_1}\right)^{n-1} .$$

From the derivations given above, it is clear how one would go about deriving from (4) the expectation of any statistic which can be defined as a sum of finite products of x_n 's. In this connection it should be noted that formula (4) does not require for its validity an infinite sequence of trials.

The Distributions of the Total Number of Errors and the Trial of the Last Error

I will now derive the probability generating function[2] of T,

i.e., the function g_{T_w} given by

$$g_{T_w}(s) = \sum_{k=0}^{\infty} P_p(T_w = k)s^k .$$

We can write T_w as

$$T_w = \sum_{m=0}^{\infty} T_{wm}$$

where T_{wm} is the number of errors made when the error probability is $a^m p$. The T_{wm} are obviously mutually independent. Thus, if $g_{T_{wm}}$ is the probability generating function of T_{wm}, we have

$$g_{T_w}(s) = \prod_{m=0}^{\infty} g_{T_{wm}}(s) . \tag{13}$$

But we can write T_{wm} as

$$T_{wm} = \sum_{j=1}^{S_{wm}} x_{wm,j}$$

where $x_{wm,j}$ is the error indicator random variable for the j^{th} trial on which the error probability is $a^m p$ and S_{wm} is the number of trials on which the error probability is $a^m p$. Since the $x_{wm,j}$ for $j = 1,2,\ldots, S_{wm}$ are identically distributed, mutually independent, and independent of S_{wm} it follows that

$$g_{T_{wm}}(s) = g_{S_{wm}}(g_{x_{wm,1}}(s)) , \tag{14}$$

where g_{S_m} and $g_{X_m,1}$ are the probability generating functions of the subscript random variables. Since S_m has the geometric distribution with parameter c (thus $g_{S_m}(u) = cu/1-(1-c)u$) and $X_m,1$ has the binomial distribution with parameters 1 and $a^m p$ (thus $g_{X_m,1}(s) = a^m ps + (1-a^m p)$), we obtain from (13) and (14)

$$(15) \qquad g_T(s) = \prod_{m=0}^{\infty} \frac{c[a^m p\, s + (1-a^m p)]}{1 - (1-c)[a^m ps + (1-a^m p)]} .$$

In particular,

$$(16) \qquad P_p(T = 0) = g_T(0) = \prod_{m=0}^{\infty} \frac{c(1-a^m p)}{1 - (1-c)(1-a^m p)} .$$

The distribution of N, the trial number of the last error in an infinite sequence of trials, is easily obtained from (16). For $n \geq 0$

$$\begin{aligned} P_p(N \leq n) &= P_p(X_{n+1} = 0,\ X_{n+2} = 0, \ldots) \\ &= \sum_{k=0}^{n} P_p(X_{n+1} = 0,\ X_{n+2} = 0, \ldots \mid Y_n = k) P_p(Y_n = k) \\ &= \sum_{k=0}^{n} P_{a^k p}(X_1 = 0,\ X_2 = 0, \ldots) \binom{n}{k} c^k (1-c)^{n-k} . \end{aligned}$$

So

$$(17) \qquad P_p(N \leq n) = \sum_{k=0}^{n} \left(\prod_{m=0}^{\infty} \frac{c(1-a^{m+k} p)}{1 - (1-c)(1-a^{m+k} p)} \right) \binom{n}{k} c^k (1-c)^{n-k} .$$

Application to Paired-Associate Learning

In a recent experiment[3] conducted by Patrick Suppes and Madeleine Schlag-Rey, each of 40 college students learned a 12 item list of paired-associates (this part of the experiment will be referred to as "the first session" or S1 below) and then learned another such list 7 days later, on the average (during the "second session" or S2). The stimuli were CVC nonsense syllables and the appropriate response to each was a press on one of three keys available before the subject. (Thus the present experiment differs from conventional paired-associate learning experiments both in the number and nature of the response alternatives available to the subject.) The order of presentation of the 12 stimuli was randomized over successive presentations. The steps of the experimental routine were as follows: A nonsense syllable was displayed before a subject. As soon as he wished (the average latency was 1.2 seconds) the subject pressed a key. A light immediately flashed over one of the keys indicating which key was correct. Four seconds later the next stimulus appeared. Each session began with three practice items after which the list to be learned was presented 25 times.

In the analyses to follow all 40 x 12 = 480 subject-items within each session are assumed to be stochastically independent and identical. Predictions of the random-trial incremental model for an infinite sequence of trials will be compared with data for the 25 trials of each experimental session. According to the model (consider Eq. 5 for $i > 25$ and the parameter values used below) the error thus introduced is quite small.

Learning was noticeably faster in S2 than in S1 (the mean total numbers of errors for the two sessions, for instance, were 3.26 and 4.37, respectively) so the model was applied separately to the data from the two sessions. I will complete most of my discussion of the data from S2 before turning to that from S1.

The a priori estimate .6667 of $\underline{p}$ was used in analyzing the data from S2. Defining the functions $\underline{a}(\underline{x},\underline{y})$ and $\underline{c}(\underline{x},\underline{y})$ by

$$\underline{a}(\underline{x},\underline{y}) = \frac{\underline{p}(\underline{x}-\underline{p})}{\underline{y}} - 1 \tag{18}$$

$$\underline{c}(\underline{x},\underline{y}) = \frac{\underline{p}}{(2-\underline{p}(\underline{x}-\underline{p})/\underline{y})\underline{x}} \tag{19}$$

it can be shown by elementary computations that

$$\underline{a}(\underline{E}(\underline{T}_{\underline{w}}), \underline{E}(\underline{u}_{\underline{w}2})) \equiv \underline{a} \tag{20}$$

and

$$\underline{c}(\underline{E}(\underline{T}_{\underline{w}}), \underline{E}(\underline{u}_{\underline{w}2})) \equiv \underline{c} \, . \tag{21}$$

Therefore the statistics

$$\hat{\underline{a}}_{\underline{w}} = \underline{a}(\overline{\underline{T}}_{\underline{w}}, \overline{\underline{u}}_{\underline{w}2}) \tag{22}$$

and

$$\hat{\underline{c}}_{\underline{w}} = \underline{c}(\overline{\underline{T}}_{\underline{w}}, \overline{\underline{u}}_{\underline{w}2}) \tag{23}$$

(where $\overline{\underline{T}}_{\underline{w}}$ and $\overline{\underline{u}}_{\underline{w}2}$ are the sample average total numbers of errors and pairs of errors) are moments estimators of $\underline{a}$ and $\underline{c}$. For the data from S2, (22) and (23) give $\hat{\underline{a}}_{\underline{w}} = .1655$ and $\hat{\underline{c}}_{\underline{w}} = .2454$. The small

value of $\hat{a}_{ww}$ is satisfying since one would not expect too much difference between two and three response paired-associate learning, and the success of the all-or-none model in predicting the course of two response paired-associate learning (see Bower, 1961) indicates that the value of a appropriate for the random-trial incremental model in such learning is very small. Estimators like $\hat{a}_{ww}$ and $\hat{c}_{ww}$ are well understood mathematically and reasonably well behaved at least in the large sample case.[4] Specifically (see Wilks, 1962, Theorem 9.3.1a, p. 260), these estimators are asymptotically normally distributed with asymptotic means a and c and asymptotic variances

$$(24)\quad \text{asy. var. } \hat{a}_{ww} = \frac{1}{n}\left(a_x^2 \operatorname{var}(T_{ww}) + a_y^2 \operatorname{var}(u_{w2}) + 2a_x a_y \operatorname{cov}(T_{ww}, u_{w2})\right)$$

and

$$(25)\quad \text{asy. var. } \hat{c}_{ww} = \frac{1}{n}\left(c_x^2 \operatorname{var}(T_{ww}) + c_y^2 \operatorname{var}(u_{w2}) + 2c_x c_y \operatorname{cov}(T_{ww}, u_{w2})\right)$$

where the partial derivatives are evaluated at $(E(T_{ww}), E(u_{w2}))$ and n is the number of subject-items. The number of subject-items in the present experiment is sufficiently large to justify consideration of asymptotic means and variances. Replacing all moments in (24) and (25) by the corresponding predictions of the model using the parameter estimates $\hat{a}_{ww} = .1655$ and $\hat{c}_{ww} = .2454$ and replacing n by 480, we obtain

$$\sqrt{\text{asy. var. } \hat{a}_{ww}} = .0297 \text{ and } \sqrt{\text{asy. var. } \hat{c}_{ww}} = .0145$$

as approximations to the asymptotic standard deviations of $\hat{a}_{ww}$ and $\hat{c}_{ww}$

for the data from S2. Certainly these values are larger than one might desire. On the other hand, they are not so large that I will feel too many qualms when I use them below.

An extensive set of comparisons of the model with the S2 data is presented in Table 1. The fit is generally excellent, though not uniformly so. The mild bimodality in the observed distribution of N_w (the same phenomenon appears in the data from the first session) is certainly not predicted by the random-trial incremental model (or any other model that I know of).[5] Also, all of the predicted $E(c_{wn})$, $n \geq 2$ are smaller than the corresponding observations (this was also observed in S1) and the effect is moderately large for $n = 4$, 5, and 6.

The reader may have noticed earlier that the theoretical expressions in the model for $P(x_{wn} = 1)$, $E(c_{wn})$, and $E(T_w^{(n)} | x_{wn} = 1)$ are exponential functions of n (see Eqs. 5, 8, and 12). It is easily seen directly from (4) that the same is true of $P(x_{wn} = 1, x_{wn+1} = 1)$ and $P(x_{wn} = 1, x_{wn+1} = 1, x_{wn+2} = 1)$. It is therefore particularly interesting to plot the logarithms of these quantities, which are linear in n, and the logarithms of the corresponding sample averages as functions of n. Such plots appear in Fig. 1.

Two analyses were made of the data from the first session, one using the estimate .6667 of p as above and the other using the estimate .6937, the proportion of the subject-items which had an error on trial 1. While the qualitative conclusions to be reached below would have been the same under the two analyses, only the second will be discussed below since it yielded a somewhat better fit. The results

Table 1

Session 2 Data and Predictions of the

Random-Trial Incremental Model

(p = .6667, a = .1655, c = .2454)

		Data	Model
$P(x_n = 1)$			
	n = 1	.6666	.6667
	2	.48	.53
	3	.43	.42
	4	.37	.34
	5	.25	.27
	6	.22	.21
	7	.18	.17
	8	.14	.13
	9	.11	.11
	10	.09	.08
	11	.08	.07
	12	.05	.05
	13	.03	.04
$P(x_n = 1, x_{n+1} = 1)$			
	n = 1	.34	.35
	2	.26	.27
	3	.23	.20
	4	.14	.16
	5	.10	.12
$P(x_n = 1, x_{n+1} = 1, x_{n+2}=1)$			
	n = 1	.19	.18
	2	.16	.14
	3	.10	.10
	4	.06	.08
$E(T)$		3.2562	*
$\sigma(T)$		2.90	2.67
$E(u_k)$			
	k = 2	1.4812	*
	3	.72	.73
	4	.35	.37
	5	.18	.19
	6	.10	.09

Table 1 (cont.)

		Data	Model
$E(c_k)$			
	$k =$ 2	1.22	1.18
	3	.97	.94
	4	.82	.74
	5	.70	.59
	6	.54	.47
$P(J = k)$			
	$k =$ 0	.33	.33
	1	.33	.31
	2	.15	.17
	3	.08	.09
	4	.05	.04
	5	.02	.02
	$\geq$ 6	.03	.02
$E(T^{(n)} \mid x_n = 1)$			
	$n =$ 1	2.91	2.59
	2	2.86	2.48
	3	2.60	2.37
	4	2.09	2.27
	5	2.06	2.17
	6	2.37	2.08
	7	2.15	1.99
	8	1.78	1.91
	9	2.10	1.83
$P(N = k)$			
	$k =$ 0	.13	.07
	1	.12	.13
	2	.09	.12
	3	.09	.10
	4	.12	.09
	5	.09	.08
	6	.06	.07
	$\geq$ 7	.29	.34

*Used to estimate parameters.

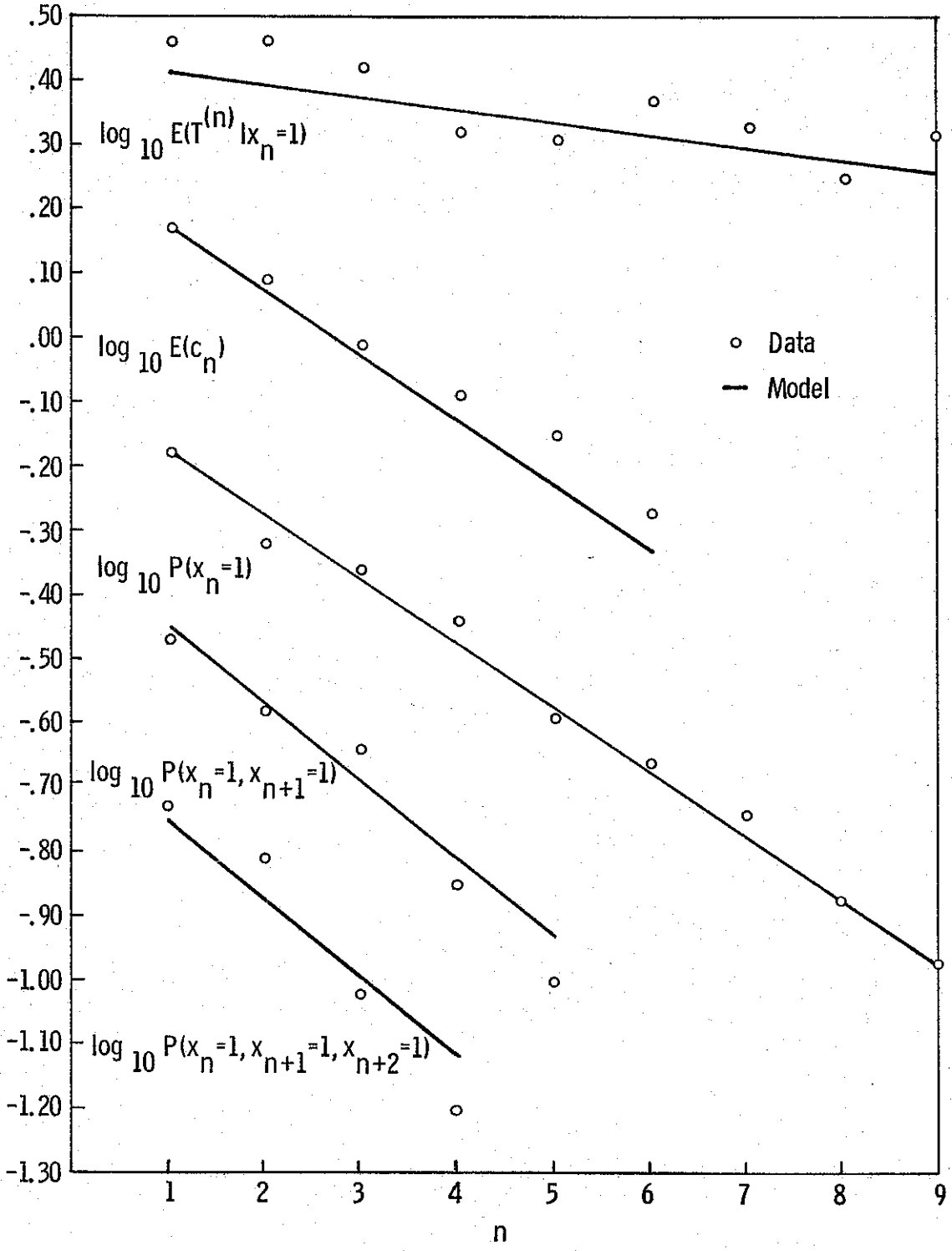

Fig. 1. $\log_{10}$ of $E(T^{(n)} | x_n = 1)$, $E(c_n)$, $P(x_n = 1)$, $P(x_n = 1, x_{n+1} = 1)$ and $P(x_n = 1, x_{n+1} = 1, x_{n+2} = 1)$: predictions of the random-trial incremental model and data for sessions 2.

are presented in Table 2.

Though the overall fit is not bad in the sense that the absolute values of the discrepancies between the model and the data are small in most cases, these discrepancies form a pronounced pattern. Examination of Table 2 shows that the model predicts too few errors, pairs of consecutive errors, and triples of consecutive errors early in learning. (Thus, as a consequence of the fact that the model predicts the total number of errors and pairs of consecutive errors throughout S1 exactly because of the way the parameters were estimated, it predicts too many errors and pairs of consecutive errors late in learning.) This generalization leads us to expect that the random-trial incremental model will overestimate the average number of errors following an error on trial $\underline{n}$ for $\underline{n}$ not too large. The denominator in the prediction will tend to be smaller than the comparable quantity in the data, and the numerator will tend to be larger than the comparable quantity in the data (see the first line in the derivation of Eq. 12). From Table 2 we see that this effect was obtained on trials 2-9.[6]

Such a pattern of errors would be expected if the random-trial incremental model were correct, but for the fact that its parameters were changing as learning proceeded. The average total number of errors in S2 was in fact a good bit smaller than the comparable statistic for S1. A natural inference is that the subjects were learning to learn gradually throughout the experiment. If such second order learning were negatively accelerated, a seemingly reasonable assumption, then its effects might have been negligible in S2. But under the further

Table 2

Session 1 Data and Predictions of the

Random-Trial Incremental Model

($\underline{p}$ = .6937, $\underline{a}$ = .1215, $\underline{c}$ = .1805)

		Data	Model
$\underline{P}(\underline{x}_{\underline{n}} = 1)$			
	$\underline{n}$ = 1	.6937	*
	2	.56	.58
	3	.50	.49
	4	.44	.41
	5	.41	.35
	6	.30	.29
	7	.29	.25
	8	.21	.21
	9	.19	.17
	10	.14	.15
	11	.14	.12
	12	.11	.10
	13	.08	.09
$\underline{P}(\underline{x}_{\underline{n}} = 1, \underline{x}_{\underline{n}+1} = 1)$			
	$\underline{n}$ = 1	.42	.40
	2	.35	.33
	3	.31	.27
	4	.24	.22
	5	.20	.18
$\underline{P}(\underline{x}_{\underline{n}} = 1, \underline{x}_{\underline{n}+1} = 1, \underline{x}_{\underline{n}+2} = 1)$			
	$\underline{n}$ = 1	.26	.23
	2	.22	.19
	3	.17	.16
	4	.13	.13
$\underline{E}(\underline{T})$		4.3749	*
$\underline{\sigma}(\underline{T})$		3.40	3.64
$\underline{E}(\underline{u}_{\underline{k}})$			
	$\underline{k}$ = 2	2.2770	*
	3	1.26	1.28
	4	.71	.73
	5	.40	.41
	6	.24	.24

Table 2 (cont.)

		Data	Model
$E(c_{wk})$	$k = 2$	1.94	1.92
	3	1.64	1.61
	4	1.44	1.36
	5	1.19	1.14
	6	.97	.96
$P(J_w = k)$	$k = 0$	.31	.31
	1	.28	.29
	2	.16	.17
	3	.09	.10
	4	.08	.06
	5	.04	.03
	≥ 6	.05	.04
$E(T_w^{(n)} \mid x_{wn} = 1)$	$n = 1$	3.86	3.68
	2	3.61	3.60
	3	3.40	3.51
	4	3.17	3.43
	5	2.90	3.36
	6	3.03	3.28
	7	2.85	3.20
	8	2.62	3.13
	9	2.66	3.06
$P(N_w = k)$	$k = 0$	.08	.05
	1	.10	.10
	2	.07	.10
	3	.06	.09
	4	.09	.08
	5	.10	.07
	6	.06	.06
	≥ 7	.42	.45

*Used to estimate parameters.

assumption that our parameter estimates for S1 were appropriate for approximately the middle of the session, and there is no reason to doubt the validity of this as a first order approximation, we would expect the model to predict too few errors early in learning and too many later in learning, just as was observed.

References

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

Bush, R. R., & Mosteller, F. Stochastic models for learning. New York: Wiley, 1955.

Bush, R. R., & Sternberg, S. H. A single operator model. In R. R. Bush & W. K. Estes (Eds.), Studies in mathematical learning theory. Stanford: Stanford Univer. Press, 1959. Pp. 204-214.

Feller, W. An introduction to probability theory and its applications Vol. 1 (2nd ed.) New York: Wiley, 1957.

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 Univer. Press, 1963. Pp. 173-187.

Wilks, S. S. Mathematical statistics. New York: Wiley, 1962.

Footnotes

[1]I am grateful to Professors R. C. Atkinson, G. H. Bower, and P. Suppes for their encouragement during this research. The preparation of this report was supported by Air Force Contract AF 49(638)-1253.

[2]The reader may consult Chs. 11 and 12 of Feller (1957) for the properties of generating functions used in this section.

[3]The data needed for the analysis presented below were generously supplied by Professor Suppes.

[4]The comment about two similar estimators on p. 182 of Norman (1963) should be ignored.

[5]A referee has pointed out that this bimodality may be indicative of a violation of the assumption of homogeneity of subject-items.

[6] $\underline{P}(\underset{\sim}{x}_n = 1)$, $\underline{n} = 1, 2, \ldots, 14$ and $\underline{E}(\underset{\sim}{T}^{(n)} | \underset{\sim}{x}_n = 1)$, $\underline{n} = 1, 2, \ldots, 8$ have been computed for the random-trial incremental model (parameter estimation as with the data from S2) for Bower's four response verbal discrimination learning experiment described in Norman (1963), and the same pattern is observed: the model predicts too few errors on the first few trials and too many on later trials while badly overestimating the average number of errors after an error on trial $\underline{n}$ for $\underline{n} = 3, 4, \ldots, 8$.

Erratum

The phrase "...and the numerator will tend to be larger than the comparable quantity in the data..." on p. 15 should be replaced by:

"...while the deficiencies of the model on the earlier and later trials will tend to cancel in the numerator..."