All-or-none theory applied to verbal discrimination learning[1]

Gordon H. Bower

Stanford University

Some recent work (Bower, 1961) has indicated the possibility of applying an all-or-none conditioning model to simplified experiments on paired-associate learning. The experimental task was arranged to circumvent response learning; Ss were told the relevant responses and were required only to learn the appropriate associations between the stimuli and available pool of responses. The model applied to these data assumed that the correct association was learned in all-or-nothing fashion, with each reinforcement providing an opportunity for learning to occur. It is assumed that before the correct association is learned S emits responses from some initial, "guessing" probability distribution over the available pool of responses.

This paper reports an extension of this theory to verbal discrimination learning (VDL). The items in the VDL task consisted of two trigrams printed on a flashcard. There were 20 such pairs constructed from 40 different trigrams. One member of each pair was arbitrarily called correct and S's task was to learn to read off the correct member as he is shown each pair in turn. In terms of the model, the important feature of this task is that S is not required to learn (or integrate) the responses. Each pair of trigrams presents a distinct pattern of stimulation and the relevant response alternatives to that pattern are embedded in it. It is assumed that the correct response

to a given pair is learned with probability c every time it is reinforced; once learned, it remains so for the duration of the short experimental session. Until an item is learned, the S presumably guesses between the two alternatives on the card, being correct a random half of the time.

The experiment described below was carried out to check the detailed quantitative fit of this model to VDL data. No experimental variables were manipulated; rather the value of this report lies in its detailed internal analysis of ordinary VDL data in terms of the expectations of the theoretical model.

Method

Materials. The items were constructed by pairing off 40 CCC trigrams selected from the 0-17 percent categories of the Witmer association values (Witmer, 1935). The resulting pairs were printed in .5 in. black, block letters on 2x3 in. flash cards. The deck of 20 cards was readily shuffled between trials. The choice of the correct responses for each S was random with the restriction that half of the correct responses were on the left and half on the right side of the printed cards.

Subjects. The Ss were 34 university students fulfilling a service requirement for an Introductory Psychology course. They were run in individual sessions. Each E tested half the Ss.

Procedure. After S was seated, he was read the following instructions: "This is an experiment on verbal learning. I have here a deck of 20 cards on each of which are printed two different trigrams. For each

card, I have selected arbitrarily one of the trigrams as the 'correct' answer. I will show you the cards one at a time; you read off what you think is the correct trigram on each card. After you respond, I'll tell you the correct answer. We'll keep going through the deck of cards until you get everything correct twice in a row. I'll shuffle the cards between trials in order to scramble their order of appearance. Try to respond within 2 or 3 seconds after I show you each card; it usually doesn't help to take longer than that." It was required that S respond to every item on every trial, even on the first trial when only guessing could occur. This procedure allowed S to control his time of exposure to the item before he responded; however, it was found that Ss did respond quickly (2 to 3 sec.) in the majority of cases. After S responded, E spelled out the correct answer, and 2 sec. later the next item was presented. The deck of cards was shuffled for 10 sec. between trials. Training continued to a criterion of two consecutive errorless trials on the whole list of items.

Results

I. General Quantitative Fit of the Model.

Before using the model to predict numerical quantities of the data, an estimate of the learning parameter c must be obtained. In the following, it is assumed that individuals and items are homogeneous; thus, there is a pool of 34x20 = 680 response sequences to account for with a single parameter. A stable estimate of c may be obtained from the mean total errors to criterion per item per subject (which was 1.99 in the data). To relate this measure to c, the following reasoning

is offered: the probability that an item is not yet learned at the beginning of the n^{th} trial is $(1 - c)^{n-1}$, which is the probability of $n - 1$ successive failures to condition the correct response to the item. The average probability of an error on trial n is then $(1 - g)(1 - c)^{n-1}$, where g is the probability of a correct response by chance guessing, and the probability that the item is learned on exactly the n^{th} trial is $c(1 - c)^{n-1}$. From the probability distribution of the trial of learning, it may be determined that it will take an average of $1/c$ trials before learning occurs. For each of these $1/c$ prelearning trials, the probability of an error is $1 - g$. Hence, the expected number of errors to criterion is $\frac{1 - g}{c}$. For our experiment $g = .50$ by assumption. Equating $.5/c$ to the observed value of 1.99 for average errors, the c estimate obtained is .251. This estimate implies that an average of four reinforcements were required before an item was learned.

Given this estimate of the learning parameter, the theory generates numerical predictions for a variety of measures (random variables) taken from the data. Derivations of the predictions have been presented elsewhere (Bower, 1961) and are not reproduced here. A first prediction is the average learning curve, i.e., probability of an error on each trial of the experiment. This curve is shown in Fig. 1 which compares

Insert Fig. 1 here

obtained and predicted values. The error probability for individual sequences supposedly starts at .50 and jumps to 0 when learning occurs. The smooth decreasing curve in Fig. 1 results from averaging across a number of sequences where the jump in response probability occurs at different trials for different sequences.

Other predictions of the model are shown in Figures 2, 3, and 4 which give the respective probability distributions of trials before the first success, total errors per sequence, and trial of last error on individual sequences. In each case the model gives a respectable approximation to the data.

Figs. 2, 3, and 4

Table 1 provides 15 comparisons of point predictions with data. More comparisons could be provided, but the sample list is sufficient to establish the point that the model provides an excellent fit to numerical details of the data.

Table 1 here

As mentioned previously, it was assumed initially that the 20 items in the list were relatively homogeneous in difficulty; that is, only one c parameter was estimated for the lot of them. It is possible to test whether this assumption of item homogeneity is realistic. Effectively one does this by comparing the observed variance in difficulty

between items with the between-item variance predicted by the model which assumes a common learning parameter for each item. Suppose the measure of difficulty for an item is the total number of errors made on it by the 34 Ss. According to the theory, the variance of the 20 items on this measure will be 34 times the variance of the total errors per sequence; the standard deviation will be $\sqrt{34}\ (1.98) = 11.6$. This predicted value compares favorably with the observed inter-item standard deviation which was 12.7. Thus, the variability between the 20 items in summed error scores is about what is expected on the basis of random sampling from the common stochastic process assumed by the model.

One question that can be raised is whether there are important initial response biases that operate during learning but they are masked in overall averages because of counterbalancing of the positions of the correct responses. The theory assumes no initial response biases in this task. There are several ways to check this assumption, each depending on a partition of the sequences into those having a correct or incorrect response on the first ("guessing") trial. One comparison is provided by the probability of a correct response on Trial 2 (after one reinforcement) following a correct or incorrect response on Trial 1. If strong initial response biases are present, then these conditional probabilities should differ markedly. However, if the assumption of no bias is correct, then the two probabilities should have a common value of $c + (1 - c)g = .626$. The observed probabilities of a correct response on Trial 2 were .642 and .612, respectively, following a correct or incorrect response on Trial 1. These measures differ negligibly and their mean (.627) is close to the predicted value. Hence, there is little indication in this test of important initial response biases.

II. Bernoulli properties of prelearning data.

One of the strong assumptions of this model is that the trial-sequence of responses to an item prior to learning may be represented as an independent Bernoulli series. That is, prior to learning an item, S's series of responses is analogous to the successive outcomes of tossing a coin. This assumption is fundamental and, fortunately, can be tested without estimating any parameters. The Bernoulli process continues until the trial of learning. However, the trial of learning is not an observable event here because S may guess correctly for several trials before he learns an item. This fact forces attention to the trials preceding (but not including) the last error in a given sequence. In the following, the responses in individual sequences prior to the last error are checked for various properties of a Bernoulli series.

(a) Stationarity. The theoretical assumption is that the probability of a correct response is constant at .50 over all trials before the last error on a given sequence. The data shown in Fig. 5 are estimates of

Fig. 5 here

this success probability pooled over all sequences. This graph was constructed by dropping sequences as their last error occurred. For example, if a sequence had its last error on Trial 6, it would enter into computations (contributing success or fail) only on Trials 1 to 5. As trials increase, the estimates become more unreliable as fewer cases are involved. Apparently the assumption of a constant .50 is reasonable since none of the estimates differ significantly from .50.

An alternative method for testing this stationarity assumption takes account of any possible differences between items and subjects in learning rates or initial response probabilities. For each sequence, the trials prior to the last error are divided into a first and second half, throwing out an odd trial in the middle if necessary. The question is whether the proportion of successes increases from the first to the second half for that sequence. According to the theory, the true success probabilities in the two halves are equal. However, in estimating these success probabilities for single sequences, the estimates will have considerable variance because, for most sequences, the estimates will be based on very few observations. Because of this variability in individual half-estimates, it is expected that for some sequences the estimate for the second half will be larger than that for the first half, while for other sequences the reverse will be true. The important prediction is that if the true success probability is constant, then the number of sequences with increasing estimates should be balanced by an equal number of sequences whose estimates decrease from the first to the second half. The logic here is the same as that involved in applying the nonparametric sign test to independent, paired observations. In analysing the data, only sequences having at least two trials in each half were used. The relevant results are shown in Table 2. This

Table 2

table shows that the number of sequences with increasing estimates (34 percent) is approximately balanced by the sequences with decreasing estimates (32 percent). Pooling respective halves for all sequences

involved, the obtained estimate of success probabilities were .492 and .521, respectively, in the first and second halves of the pre-learning trials. Both of these estimates are close to the assumed value of .50 and the small difference between them is less than the standard error of the difference.

If attention is restricted to those sequences having a specified number of observations in each half, then exact prediction can be derived from the binomial distribution for the respective percentages of sequences that increase, decrease, or stay the same in success probabilities from the first to second half. For these purposes, sequences with exactly two observations in each half were employed; there were 101 of these from the original pool of 680 sequences. From considerations of the binomial distribution it may be shown that the probability of getting more successes in the second than in the first half is $p^2q^2 + 2pq(p^2 + q^2)$, where $p = 1 - q$ is the probability of a success. Letting $p = .50$ the predictions in Table 3 are obtained

Table 3

and are compared with data. These data indicate that the probability of a correct response is constant .50 prior to the last error.

(b) Independance. Another important feature of a Bernoulli series is that successive events are statistically independent. In a coin tossing experiment the probability of a tail is independent of the preceding outcomes of tosses; similarly, here it is assumed that prior to the

last error in the sequence the probability of a correct response is the same following a correct or incorrect response on the preceding trial. Independence is a very strong assumption about response sequences (response-response dependencies are the rule in psychology) and one is satisfied if the assumption is even approximately correct. There is no necessary entailment between stationarity and independence aspects of an event series, so statistical tests of these assumptions are independent. In testing for independence the frequency of occurrence of the four possible one-trial transitions between correct and incorrect responses were tabulated. The last error was included (as a final trial) in this tabulation since no bias is thus introduced when comparing transition probabilities. The resulting 2x2 table of frequencies yielded a chi-square value of 3.31 ($df = 1$, $P > .05$) which indicates that the independence assumption cannot be rejected. The power of this chi-square test is substantial since it was based on a total frequency of 1719 transitions. The estimated probability of a correct response following a correct response was slightly higher than that following an incorrect response. This difference reflected the fact that on a few occasions Ss gave a fairly long string of correct responses and then made a final error on an item. Apparently, the frequency of such success runs was a little larger than expectations from chance guessing.

(c) Binomial distribution. Another feature of Bernoulli random variables is that their sums have a binomial distribution; that is, the probability of exactly x successes in n observations will be

$$b(x;n) = \binom{n}{x} p^x (1-p)^{n-x} \qquad (1)$$

where p is the probability of a success in each observation. In the solid curves of Figures 6 and 7 are shown the obtained probability distributions of the number of successes in blocks of 4 and 6 prelearning trials, respectively. The dashed curves give the probability values calculated from Equation 1 with the assumption that $p = .50$ and n equal to 4 and 6, respectively, in the two figures. The mean and mode of the empirical distributions are properly located and they have the proper symmetry about the mode. The notable deviation in both cases is that the data exhibit slightly more long runs (of errors and successes) than was predicted. These same slight discrepancies were evident in the analysis of the independence of successive responses.

Figs. 6 and 7

The preceding analyses confirm the first critical assumption of the theory; namely, that the sequence of responses to an item prior to the last error may be represented as a series of independent Bernoulli trials.

III. Assumption of constant probability of learning.

The second central assumption of the theory is that the probability of learning an unlearned item remains constant over trials. That is, if S has failed to learn an item in n previous trials, his probability of learning it on the next trial is still the same constant, c. This assumption is logically independent of the stationarity assumption in the sense that tractable models could be devised with stationarity

but, say, with c increasing with practice. As seen before, stationarity can be checked in the data without estimating learning parameters, so those tests were independent of assumptions about the learning rate. However, tests of assumptions about c necessarily assume stationarity. With the preceding positive results on stationarity, it will be considered as a given in testing the hypothesis that c remains constant over trials.

One simple and direct test of the constant-c assumption is the comparison of obtained and predicted probability distributions of the trial of the last error on single sequences, which was shown in Fig. 4. The trial of the last error is a sensitive test statistic because it is primarily a function of c, the probability of learning per trial. If the probability of a correct guess, g, is high, then the average trial of the last error occurs appreciably sooner than the hypothetical trial of conditioning (compare $1/c = 4.00$ with our observed mean trial of last error = 3.32) because S may guess correctly for several trials before the trial of learning. To this extent, then, the trial of the last error depends on g as well as the constant-c assumption. Since the predicted distribution in Fig. 4 corresponds closely to the obtained distribution, the constant-c assumption is supported. It is clear that if c had been changing with practice, then the predicted distribution (assuming a constant c) would have been quite discrepant from the data.

A second test of the constant-c assumption is obtained by considering the conditional probability of a success on the next trial given n consecutive failures from the beginning of practice. If the constant-c assumption is correct, then this conditional probability should be

$c + (1 - c)g = .625$ regardless of the number of immediately prior errors in the sequence. The relevant results are shown in Table 3. Starting with 680 sequences on Trial 1, the second column gives the number of sequences having the number of consecutive errors given in the corresponding row of the first column. Of these cases, the obtained and predicted number of successes on the next trial are shown in the last two columns of Table 4. The predicted numbers are the nearest integer values to

Table 4

62.6 percent of the respective observed number of cases. None of the predictions are very far from the observed numbers, so the constant-c assumption is again confirmed. It may be noted also that most of the predicted numbers are slightly larger than the observed numbers. A c estimate of .21 would give a slightly better fit to the observed numbers in Table 4.

A third method for testing the assumption of a constant learning probability is to consider the average number of errors to criterion following an error that occurs on some trial n. The theory predicts that the average number of subsequent errors will be a constant, $(1 - c)(1 - g)/c$, regardless of where the error occurs in training. When an error occurs, we know that the item is not yet learned which, of course, is the way it started. Hence, for predicting future behavior to that item, we could just as well set the clock back to Trial 1. An analogous statement in terms of a coin-tossing experiment would be

that the failure of a head to appear during the first N tosses does not affect the average number of subsequent tosses required to get the first head. The relevant results here are shown in Fig. 8 which gives

Fig. 8

the mean number of subsequent errors to criterion for those sequences having an error on Trial 1, for those with an error on Trial 2, and so on up to Trial 6. The analysis was not carried beyond Trial 6 since the number of cases was becoming small as progressively more items were learned. It should be noted that if a given sequence had errors on, say, Trials 1 and 4, it would be involved in the computations of means for both these points of Fig. 8 (although contributing one less error in the latter case). The dashed horizontal line is the constancy prediction of the all-or-none theory and the data are in good accord with this prediction. The predicted value is $(1 - c)(1 - g)/c = 1.49$ while the average of the six data points is 1.48. The smooth decreasing curve in Fig. 8 gives the predictions of the linear model (Estes, 1959) which assumes that the error probability for single sequences decreases by a constant fraction, $1 - \theta = .749$, with each reinforcement. According to this theory, the number of subsequent errors should decrease with the number of reinforcements prior to any given error. Obviously this incremental theory is not descriptive of the data.

Discussion

The theory that has been tested makes only two critical assumptions: first, that response probabilities to single items have only two values, some initial value, g, and a terminal value of unity; and second, that the probability of a shift from the initial to the terminal value is a fixed parameter, c, for every reinforced trial. Both assumptions have received substantial support from the preceding analyses of verbal discrimination data.

The results are somewhat at variance with ordinary conceptions of learning as involving the accumulation of increments in associative strength. An incremental theory combined with a threshold concept, such as Hull (1943) employed, would be consistent with these data under the following restrictive assumptions: (a) the range of oscillation of reaction potential is effectively zero--this assumption implying sudden jumps from chance to perfect responding, and (b) the initial habit strengths or learning rates for various items and Ss are distributed in some unique manner so that the trial at which individual habits first exceed the threshold value is distributed as $(1 - c)^{n-1}c$. Obviously, with these two restrictive assumptions Hull's theory becomes virtually identical with the all-or-none theory and the choice between them is a matter of convenience.

At a different level of analysis incremental theorists might inquire into possible situational factors in the VDL task that resulted in data approximating the all-or-none model. First, the VDL task circumvents response learning or response integration, which may well

be the primary factor involved in the gradual learning typically found in, say, paired-associate experiments. Second, the relatively rapid rate of learning may have masked normally prominent effects of gradual strengthening of responses. The mean trial of the last error on individual sequences was only slightly greater than 3. Since the first trial was an assured guess, this means that, on the average, there were only two trials per sequence on which suboritorial strengthening of responses could be in evidence. Third, with the high initial success rate of .50 it may be difficult to show slight improvements due to partial learning of an item. Thus, it is possible that if the experiment were run with more than two alternatives per card, partial learning might be revealed by S eventually confining his responses to a small subset of the alternatives on a card. Experimental tests of these suggestions are currently under way.

Summary

Thirty-four Ss learned 20 items in a verbal discrimination task. The data showed that (a) response sequences prior to the last error could be represented as an independent Bernoulli series, and (b) the probability of learning an unlearned item was constant over trials. A variety of numerical predictions compared favorably with the data.

Figure Captions

Fig. 1. Average probability of an error on successive practice trials.

Fig. 2. Probability distribution of the number of trials before the first success on single sequences.

Fig. 3. Probability distribution of the total number of errors to criterion on single sequences.

Fig. 4. Probability distribution of the trial of the last failure, lumping together adjacent values. The point at 0 includes only sequences having no errors throughout training.

Fig. 5. Probability of a correct response over trials before the last error. Successive points are based on fewer observations since sequences are dropped from computations as their last error occurs.

Fig. 6. Probability distribution of number of successes in blocks of 4 prelearning trials.

Fig. 7. Probability distribution of number of successes in blocks of 6 prelearning trials.

Fig. 8. Average errors to criterion following an error occurring on trials 1 to 6. See text for explanation.

References

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

2. Estes, W. K. The statistical approach to learning theory. In Koch (ed.): Psychology: A Study of a Science, Study I, vol. 2. New York, McGraw-Hill, 1959.

3. Hull, C. L. Principles of Behavior. New York: Appleton-Century-Crofts, 1943.

4. Witmer, L. R. The association value of three-place consonant syllables. J. Genet. Psychol., 1935, 47, 337-360.

Footnotes

1. Miss Ann Newton assisted the writer in this experiment. The research was supported by a grant M-3849 from the National Institutes of Mental Health.

Table 1

Comparison of observed and predicted statistics.

Statistic	Observed	Predicted
Total errors to criterion	1.99	---
Standard Deviation	1.94	1.98
Errors before first success	.80	.80
Standard Deviation	1.15	1.13
Errors before second success	1.34	1.29
Standard Deviation	1.48	1.51
Trial of last error	3.32	3.20
Standard Deviation	3.22	3.45
Alternations of success and fail	1.99	1.99
Total runs of errors	1.24	1.24
Runs of exactly 1 error	.81	.80
Runs of exactly 2 errors	.26	.29
Runs of exactly 3 errors	.10	.11
Number of joint errors 1 trial apart	.74	.74
2 trials apart	.57	.56
3 trials apart	.43	.43

Table 2

tationarity test: see text for explanation

Success probability from first to second half	Percentages of all sequences in category
Increases	.34
Stay same	.34
Decreases	.32
Average percentage success	
first half	.492
second half	.521
overall	.506

Table 3

Exact prediction of percentages
in each category for
sequences with two observations in each half.

Relation of second to first half	Predicted percentage	Obtained percentage
Increases	.31	.30
Decreases	.31	.32
Stays the same	.38	.38

Table 4

Number of successes following n consecutive errors from start of training.

Number of prior errors	Number of cases	Observed number of successes	Predicted number of successes
0	680	340	340
1	340	204	212
2	136	84	85
3	52	29	32
4	23	13	14
5	10	7	6

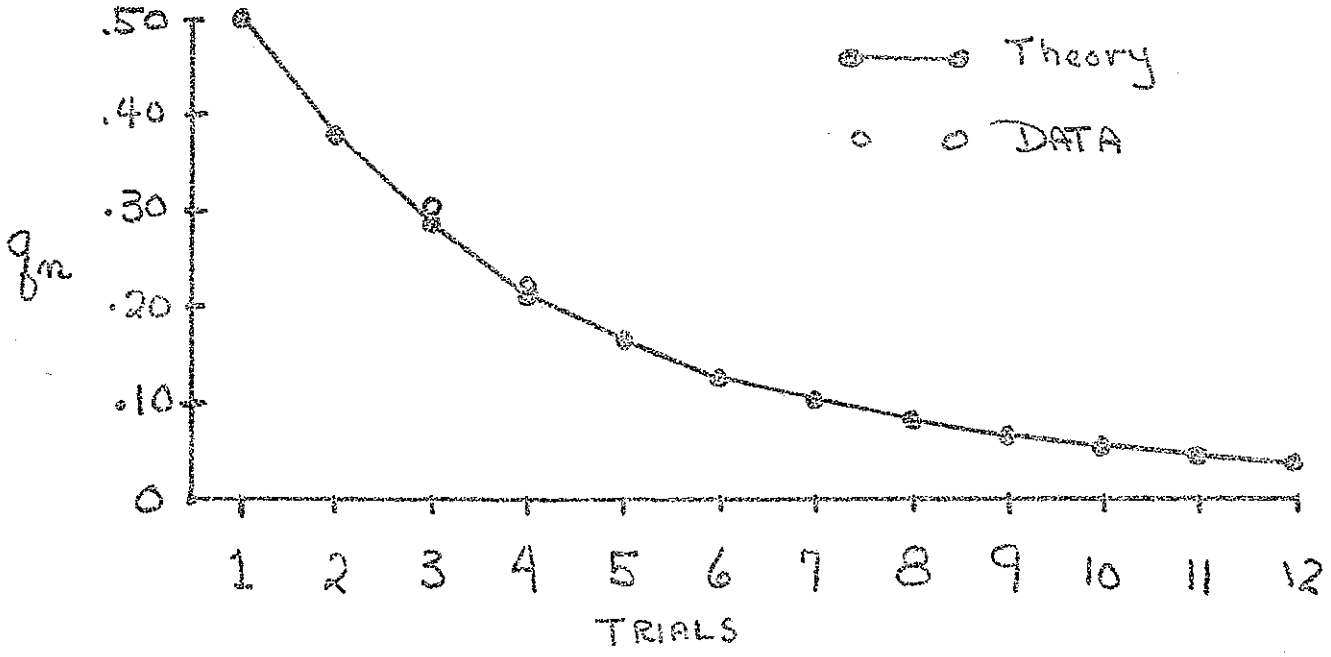

Fig. 1. Average probability of an error on successive practice trials.

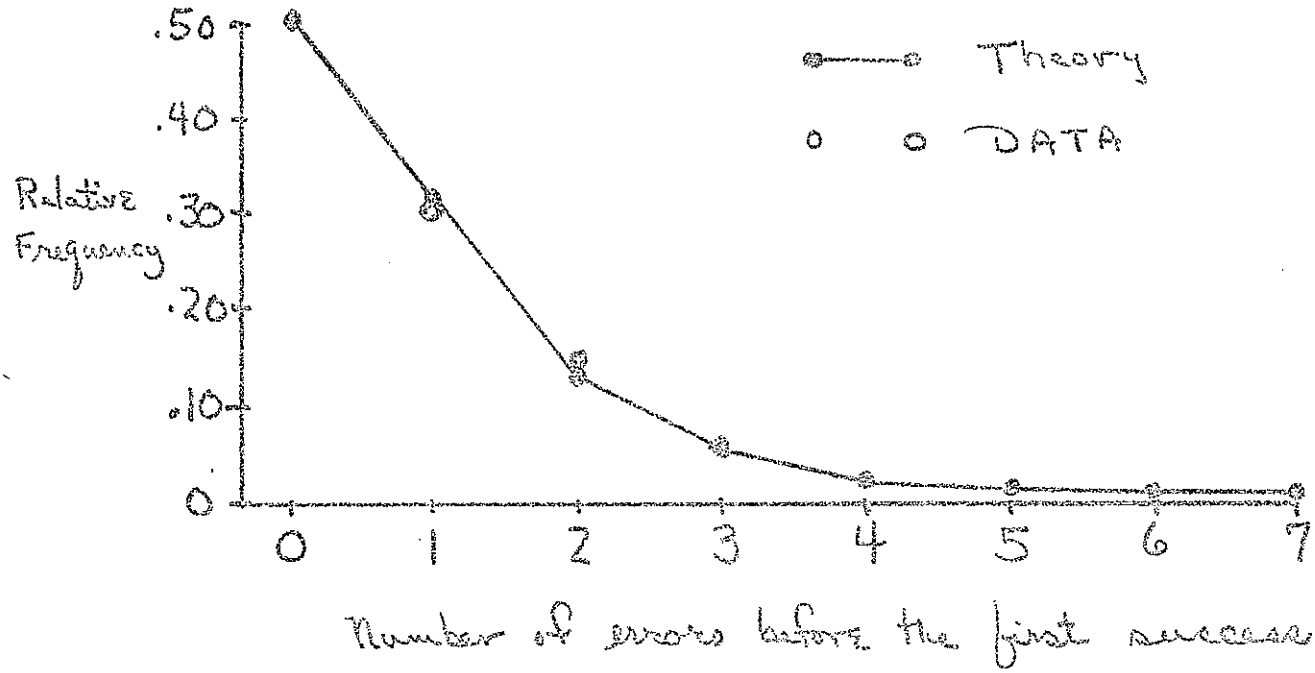

Fig. 2. Probability distribution of the number of errors before the first success.

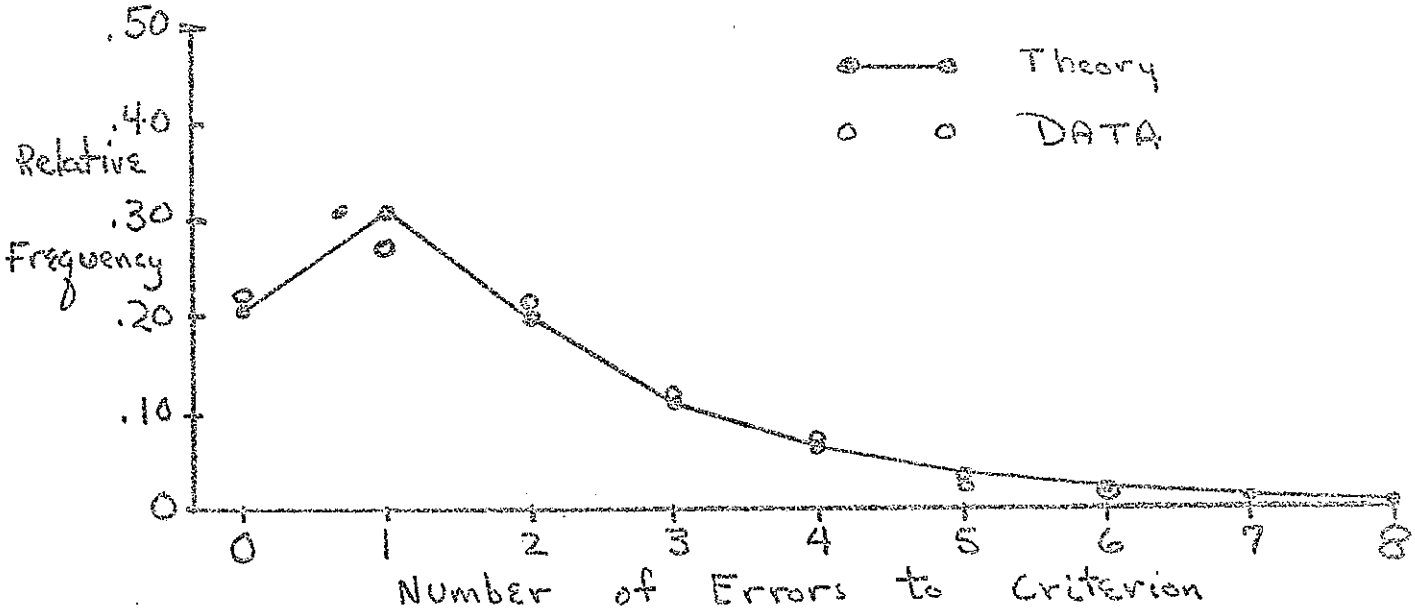

Fig. 3. Probability distribution of the total number of errors per item per subject.

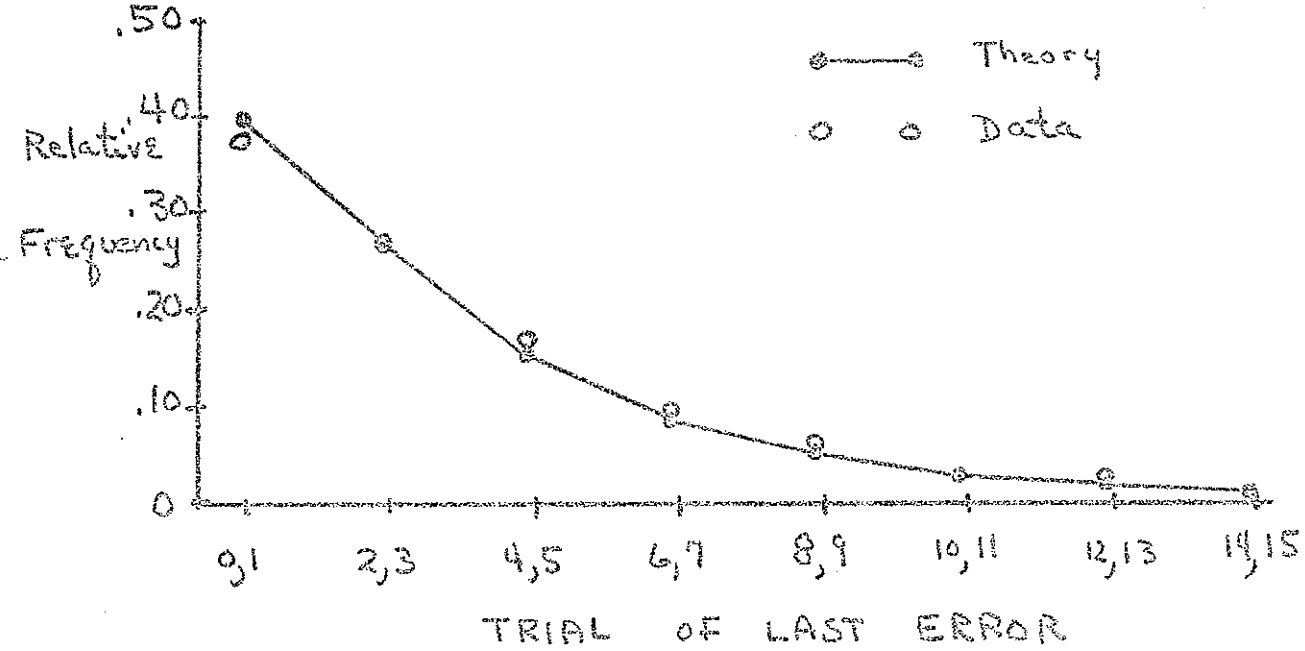

Fig. 4: Probability distribution of the trial of the last error, lumping together adjacent trials. The point at 0 includes only sequences with no errors throughout training.

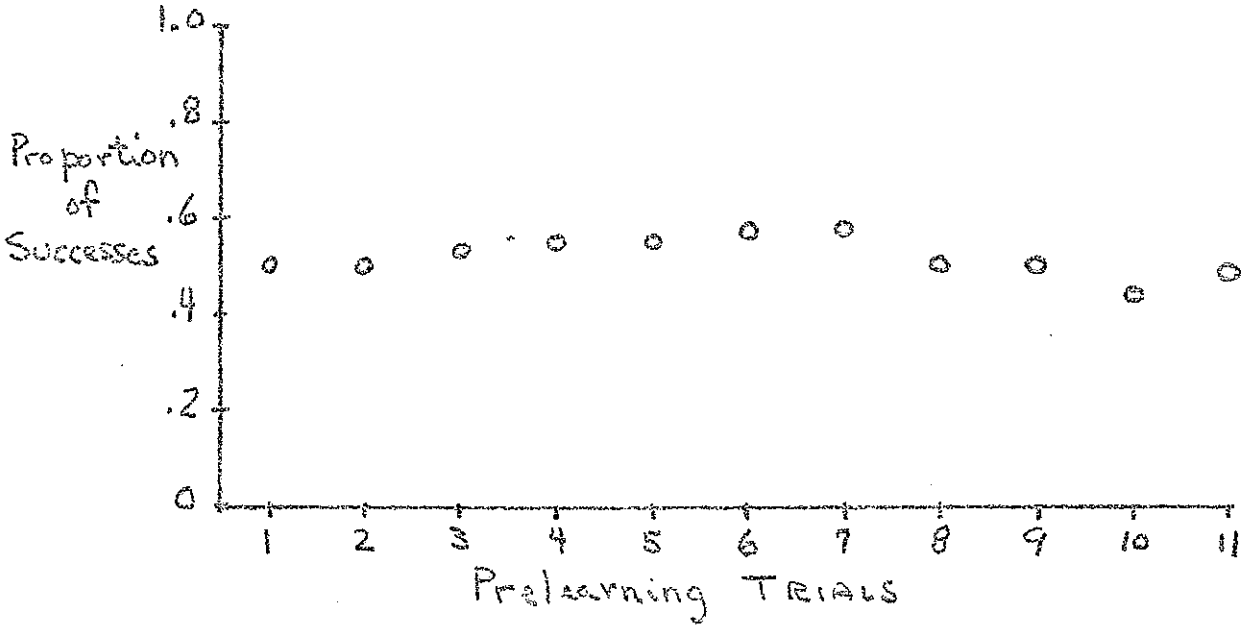

Fig. 5. Probability of a correct response over trials before the last error. Successive points are based on fewer observations since sequences are dropped from computations as their last error occurs.

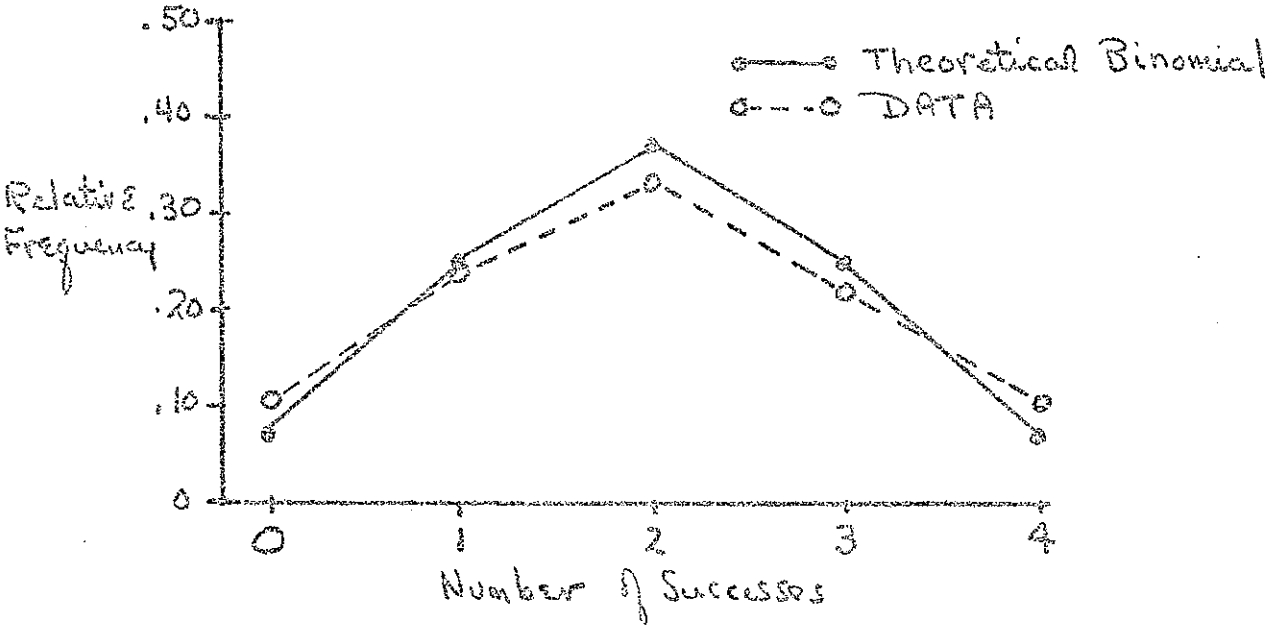

Fig. 6. Probability distribution of the number of successes in blocks of 4 prelearning trials.

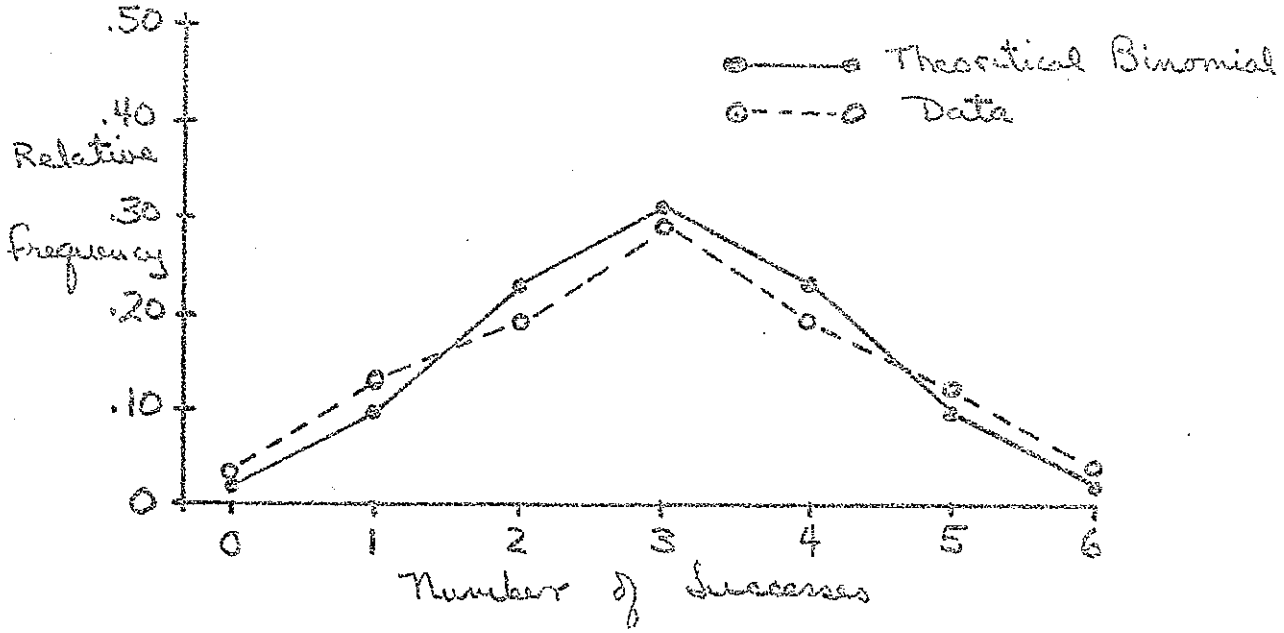

Fig. 7. Probability distribution of the number of successes in blocks of six preliminary trials.

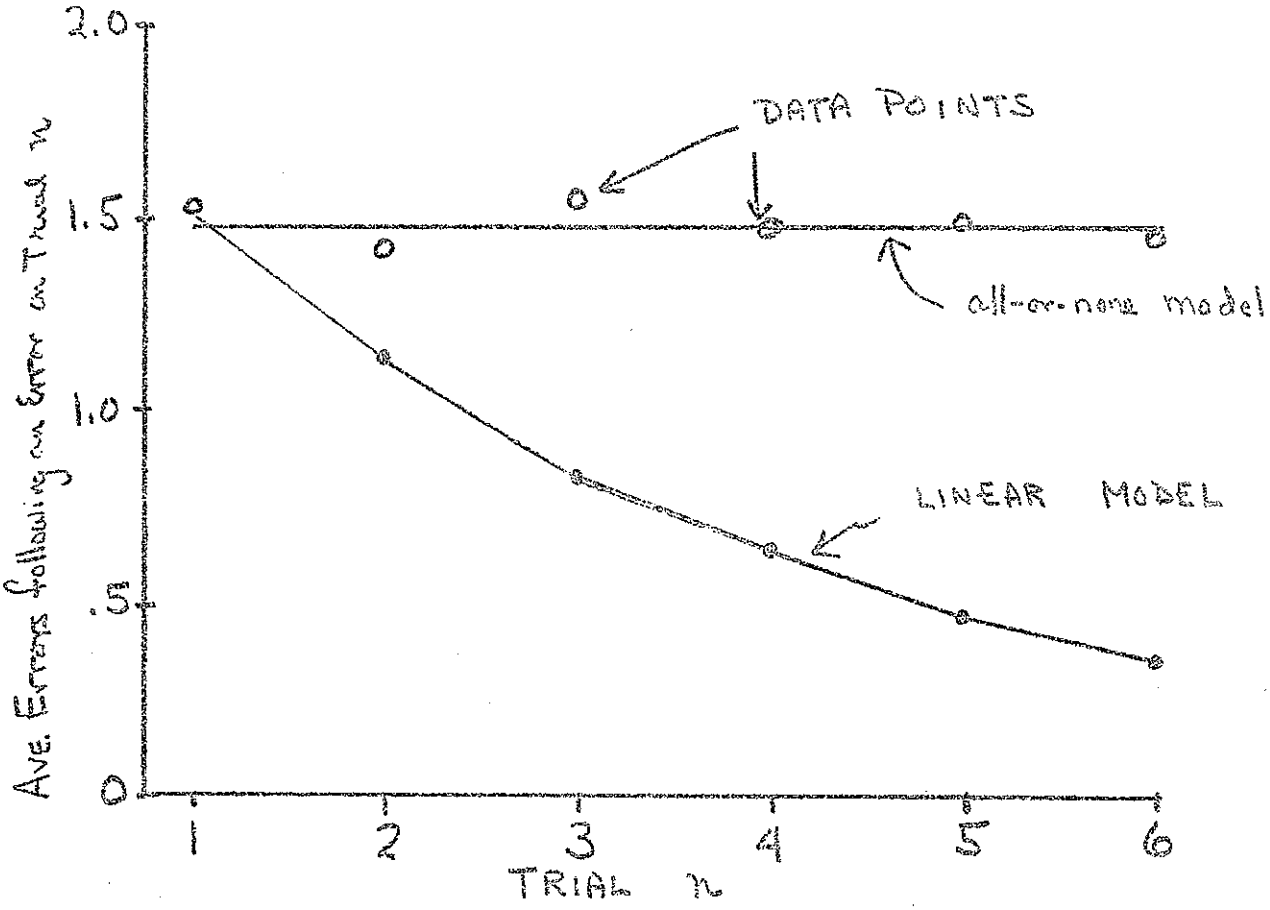

Fig. 8. Average errors to criterion following an error occurring on Trials 1 to 6. See text for explanation.