# All-or-None Subprocesses in the Learning of Complex Sequences[1]

ROBERT A. BJORK

*University of Michigan, Ann Arbor, Michigan 48104*

This paper reports a study designed to investigate whether the all-or-none conception of the learning process can be extended to a learning task more complex than conditioning or simple verbal association. The experimental task is to learn numerical sequences by anticipating each new member of the sequences. Although the obtained sequence learning appears very complex, it proves to be analyzable into constituent all-or-none subprocesses.

The success of specific models for conditioning and verbal association that embody the all-or-none conception of the learning process has led to speculation (Estes, 1964; Restle, 1964) that it might be profitable to extend the all-or-none conception to more complex learning. Although the simplest all-or-none model, the one-element model, which assumes a single all-or-none learning step is obviously inadequate for situations more complex than conditioning or simple association, it may be the case that complex learning consists of a number of successive and/or simultaneous all-or-none processes. There is some evidence for such a position. Bower and Theios (1964) describe avoidance behavior well with a model based on the assumption that learning the avoidance response involves two successive all-or-none subprocesses. Also, Restle (1964) interprets difficult paired-associate learning in terms of a multistage model in which the stages are all-or-none processes corresponding to association, stimulus discrimination, and response discrimination.

This paper reports a study designed to investigate whether the all-or-none conception can be extended to sequence learning, a learning task considerably more complicated than simple association. The general format of sequence learning is as

follows: the subject is presented the first member of a sequence, attempts to predict the second member, is presented the second member of the sequence, attempts to predict the third member, is presented the third member of the sequence, and so on until he achieves an arbitrary learning criterion of $N$ successive correct anticipations. There are two main reasons why the learning of numerical sequences is an appropriate task for this study:

1. As a learning task it is considerably more complex than those tasks which typically exhibit all-or-none learning. In simple associative tasks the stimulus and associated "correct" response are constant. In sequence learning the subject must abstract from the progressing sequence constant relationships upon which to base his prediction of the next sequence member. For example, if the sequence is 0, 2, 1, 3, 2,..., he must learn that whenever the sequence is incremented by 2 it will then be decremented by 1. In addition, most sequences involve more than one abstract association; in order to completely learn the sequence the subject must abstract several relationships concurrently.

2. Although the task is considerably more complex, the response protocols generated are formally similar to those resulting from simple association experiments. The subject's responses for a particular sequence can be represented as a trial-by-trial string of errors and successes yielding protocols essentially equivalent to paired-associate or conditioning protocols. This protocol equivalence means that differences and similarities between sequence learning and simple association are maximally apparent, since the sequence data can be analyzed in a fashion parallel to the standard analyses performed on simple association data.

## METHOD

Fifty-nine introductory psychology students at Stanford University participated in this experiment in fulfillment of a course requirement. They were each required to learn 12 numerical sequences which varied considerably in their complexity. All subjects were given the 12 sequences in the same order.

*Apparatus.* A black 3 ft. square of plywood was supported vertically on a table, separating the subject and experimenter. In front of the subject, mounted on the plywood, was a white circle of 12 in. diameter. The 10 digits, 0, 1,..., 9, were printed in black at equal intervals inside the circumference of the circle. Mounted just outside the circle next to each digit was a small light. Any particular digit could be designated by turning on its adjacent light.

The experimenter could control the digit presented on a particular trial by means of a ten-contact rotary switch wired to correspond to the ten-digit display. As the switch was turned from one number to another, all numbers in between flashed momentarily as their contacts were touched. The direction of this series of brief flashes enabled the subject to tell whether addition or subtraction was being performed. The nature of the apparatus meant that any sequence member, $N$, appeared as $N$ (modulo 10) to the subjects. Thus, if two successive members of a sequence were 8 and 12, the subjects saw 8 and 2.

*Instructions.* Subjects were told that their task would be to discover the scheme or system underlying each numerical sequence. They were also told that the sequences were arithmetic in that each new member of a sequence would be obtained by adding or subtracting some integer to or from the last member of the sequence; it was emphasized that multiplication or division were never involved. Two illustrative sequences were presented and explained to familiarize the subject with the procedure.

*Procedure.* Subjects were presented the members of a given sequence one by one. After each member was presented the subjects were required to attempt to predict the next member of the sequence. At any one time only the current member of a sequence was displayed to the subject; that is, with the presentation of each new member of the sequence the preceding member was removed.

Subjects responded vocally; the experimenter recorded each response and presented the next member of the sequence. When a subject anticipated five consecutive sequence members correctly, he was told that he had solved the sequence; there was a short interval; and then the experimenter presented the first member of the next sequence. If a subject had not started a criterion run of five correct by the twenty-fifth trial, he was told that that was the end of the sequence, and after a short interval, the next sequence was started.

*Nature of the sequences.* The experimental sequences, shown in Fig. 1, are generated by operations of two basic kinds. The simpler of the two operations is adding (subtracting) a constant integer. For example, sequence 4: 1, 4, 7, 10,..., is constructed by adding 3 on every trial. By combining several such operations more complex sequences can be generated; sequence 3: 1, 5, 7, 11, 13,..., is constructed by alternately adding 4 and adding 2.

The second, and more complicated of the two operations, is an advancing, rather than constant, increment or decrement to the sequence. For example, sequence 1: 1, 2, 4, 7, 11,..., is formed by adding 1, adding 2, adding 3, etc. Again, more complex sequences can be generated by operations of this kind by using more than one such operation in a single sequence (e.g., sequences 2 and 11).

Finally, the two operations can be mixed to construct sequences as in sequence 7.

## RESULTS

The learning curves obtained for the 12 sequences are shown in Fig. 1. These learning curves, to say nothing about more sensitive characteristics of the data, appear very complex. They show numerous irregularities which seem unlikely to be predicted by any simple model.

In spite of the general complexity of the sequence data, the possibility is not excluded that subprocesses, i.e., components of the learning defined in some manner, might exhibit all-or-none learning. A process generated by several simple, but interlaced, stochastic subprocesses could appear complex. Even if the component subprocesses are the same, their probabilistic nature and joint generation of the learning process could yield very complex performance.

A subprocess analysis requires a partition of the data made according to some rational breakdown of the learning into components. One reasonable partition is in terms of the constituent subrules shown in Fig. 1. In order to learn any of these

sequences a subject must learn the component subrule(s). The learning data from
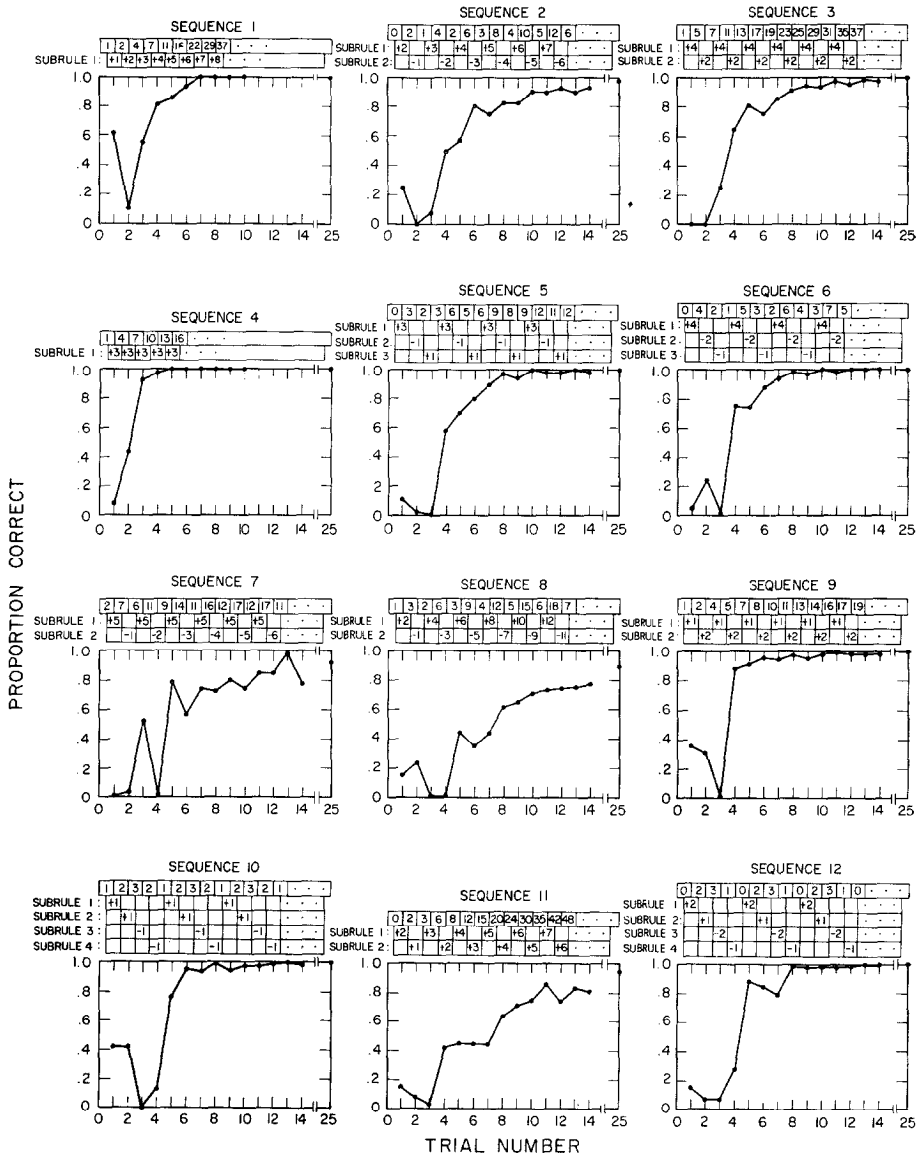this experiment are recorded in such a form that the learning of these subrules can be



FIG. 1. Learning curves for sequences 1–12. Above each learning curve are the actual
sequence members together with the subrules used to generate the sequence.

studied directly. If the trials on which a particular subrule (e.g., the subrule "add 4" in sequence 6) is applied (trials 1, 4, 7, etc.) are extracted from total protocol for the sequence, all standard learning analyses may be applied to this subset of the data.

The learning curves for any particular subrule can be obtained directly from Fig. 1 by plotting the proportion correct as a function of only those trials to which the subrule applies. When the learning curves in Fig. 1 are broken, in this manner, into subrule learning curves, striking order emerges. One such transformation is shown in Fig. 2; when the single complex learning curve for sequence 6 is broken down, three very orderly subrule learning curves emerge.
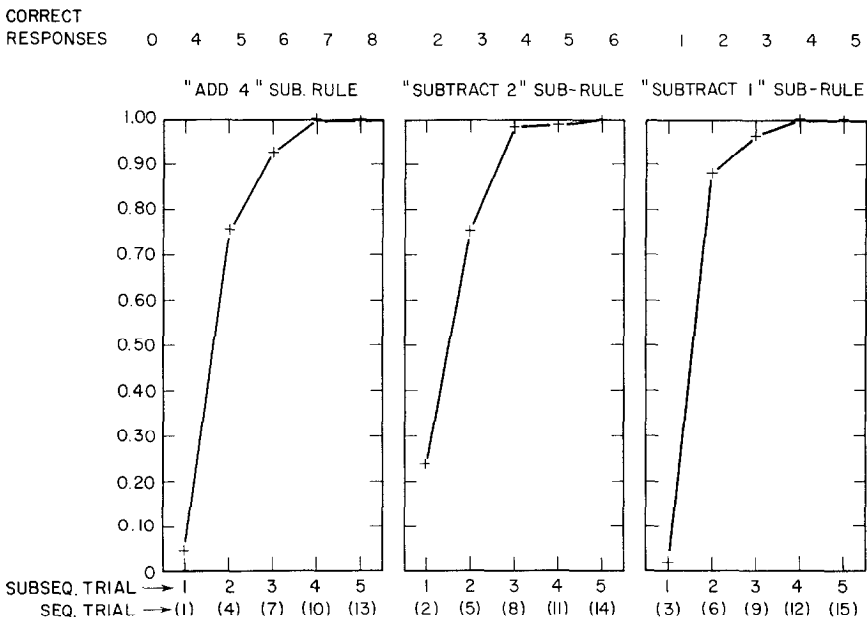


FIG. 2.    Learning curves for the three component subrules of sequence 6.

Their orderliness is encouraging, but it is not clear what kind of orderliness is involved; it is well known that such classic negatively accelerated learning curves can come from an intrinsically all-or-none process, an intrinsically incremental process, or from combinations of these.

In order to subject this issue to a detailed analysis the twelve sequences were partitioned into subsequences corresponding to the component subrules. The response protocols for these subsequences were printed on IBM cards so that the learning of each subrule could be investigated separately. For any given sequence the data from subjects who did not learn were discarded. All subjects reached criterion on sequences

1, 3, 4, 5, 6, 9, 10, 12. One, eight, seven, and four subjects failed to solve sequences 2, 7, 8, and 11, respectively.

*Organization of the Subsequence Data*

The number of subrules in any particular sequence varies from one to four and there are 28 total subrules in Fig. 1. A specific subrule and its corresponding subsequence of trials will henceforth be designated by the number of the sequence to which the subrule belongs followed by a number indicating its order of occurrence in the sequence. For example, subrule 5-2 is the second subrule ("subtract 1") of sequence 5.

Any particular subsequence can also be categorized in terms of the kind of operation required, either a constant increment (decrement) or an advancing increment (decrement), and the *period* of application of the subrule, either every trial or every second trial ... or every *n*th trial. Table 1 shows that when the 28 subrules are categorized in this manner they fall into six groups: constant subrules with periods 1, 2, 3, 4, and advancing subrules with periods 1, 2. Any such group will be referred to by a letter indicating the kind of operation involved followed by a number indicating

TABLE 1

Subrule Categorization[a]

|  | Type of operation involved in the subrule | |
| :---: | :---: | :---: |
| Period of Application | Constant increment (decrement) | Advancing increment (decrement) |
| 1 | Group C-1: 4-1. | Group A-1: 1-1. |
| 2 | Group C-2: 3-1, 2; 7-1; 9-1, 2. | Group A-2: 2-1, 2; 7-2; 8-1, 2; 11-1, 2. |
| 3 | Group C-3: 5-1, 2, 3; 6-1, 2, 3. | Group A-3: — |
| 4 | Group C-4: 10-1, 2, 3, 4; 12-1, 2, 3, 4. | Group A-4: — |

[a] The subrules are grouped according to their type of operation and their period of application in the sequence.

the period of application. Thus, groups C-3 and A-2 designate the constant subrules with period 3 and the advancing subrules with period 2, respectively.

*An All-or-None Model for Subrule Learning*

The all-or-none model applied to the subsequence represents a slight modification of the simple one-element model. Its principal substantive assumptions are as follows:

1. On trial 1 the subject is in an unlearned state and has probability $g_1$ of being correct. On the first trial of any given subsequence the subject has no information about the subrule at all; he has never seen an occurrence of the subrule in the current sequence. It would be a mistake, however, to think that the subject is guessing in the normal sense. Instead, he is usually operating on hypotheses resulting from his having observed the first occurrence of other subrules. His probability of being correct depends entirely on the fortuitous interaction of prior occurrences of other subrules and the nature of the present subrule; it may vary from zero to virtually unity, but it indicates nothing about the subject's degree of learning on the subrule. Hence, $g_1$ for any subrule is just taken to be the observed proportion of correct responses on the first subsequence trial.

2. Upon the occurrence of a reinforcement of any subrule the subject has probability $c$ of learning, in which case he will make no more errors on the subsequence, and has probability $1-c$ of not learning.

The all-or-none conception can only be tested after the first reinforcement of a particular subrule. In the case of advancing subrules this means that the subsequence protocols must start with the second subsequence trial, ignoring the first completely, because two operations must occur before the subject can detect an advancing subrule. Consider sequence 1. After the first trial the subject has only seen "add 1." The first reinforcement of the subrule, "add 1 more each trial," comes only after the second trial when the subject also has seen "add 2."

3. If the subject does not learn, he has some probability $g$ of guessing correctly on the next trial. In general, $g$ will not equal $g_1$. It is this inequality which differentiates this model from the one-element model in which $g = g_1$.

*Fit of the All-or-None Model to the Subsequence Data*

It is not manageable to exhibit the fit of the model to the obtained learning on each of the 28 subrules. Instead the model will be tested against the subrule data grouped as in Table 1. This is possible because the learning on subrules of the same type, i.e., subrules with the same kind of operation and the same period, is similar. Also, the fit to the grouped data is typical, other than being more stable, of the fit to the individual subrule data.

There is one distinct exception to the rule that the obtained learning on subrules

of the same type is similar. The subrules in group A-2 of Table 1 fall into two difficulty classes, A-2E (easy) and A-2H (hard) containing subrules 2-1, 2-2, 7-2, and 8-1, 8-2, 11-1, 11-2, respectively. Thus, the model will be tested against the subsequence data pooled into seven groups: C-1, C-2, C-3, C-4, A-1, A-2E, and A-2H.

Many features of the subrule learning can be utilized to test the model (Bower, 1961). Three features which embody most of the distinguishing characteristics of the all-or-none model are (a) the form of the learning curve, (b) the distribution of total errors, and (c) the distribution of the trial of last error.

In order to derive predictions from the model for these features of the learning, It is necessary to estimate $c$ and $g$. The guessing probability, $g$, can be estimated from the mean proportion correct after the first trial and before the trial of last error. Instead of estimating $g$ separately for each group of subrules, a value of .25, which represents the approximate mean for all the subsequences will be used. The probability of learning on a given trial, $c$, can be estimated from the mean total errors,

$$M_T = \frac{1 - g_1 - g(1 - c)}{c}.$$

When this equation is solved for $c$, an estimate of $c$ for each subrule group is obtained. That is,

$$c = \frac{1 - g}{M_T + g_1 - g},$$

where $M_T$ and $g_1$ are observed values and $g = .25$.

The predicted vs observed learning curves and the predicted vs observed distributions of total errors are presented, for the constant subrules, in Fig. 3 and, for the advancing subrules, in Fig. 4. With the exception of groups C-1 and C-2, the fit of the model would be difficult to improve upon. The predicted vs observed distributions of the trial of last error are very similar to the fit of the total error distributions and are omitted.

In groups C-1 and C-2, where the fit is not good, there are fairly obvious experimental reasons why the data are not typical. Group C-1 contains only subrule 4-1, the only instance of a constant subrule with period 1. Subrule 4-1 is the subrule "add 3" in sequence 4: 1, 4, 7, 10,..., which is so simple subjects expressed "it couldn't be that easy," and predicted in accordance with some more complex hypothesis. To accurately predict such behavior a model would have to be somewhat more idiopathic in nature than those considered in this paper. Group C-2 is atypical because the nature of sequences 3 and 9 led the subjects, during the early trials, to suppose that a subrule of the advancing sort was generating the sequence.

The close correspondence between the predictions derived from an all-or-none model of the subrule learning and the obtained subrule data virtually rules out the
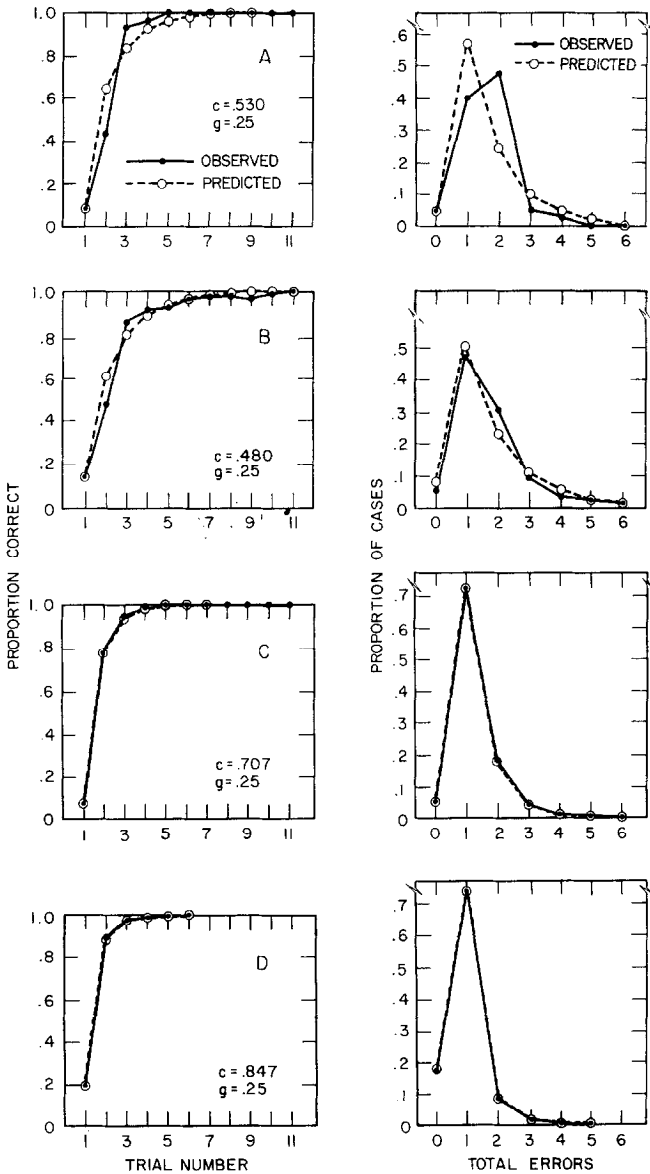
FIG. 3.   Subsequence analysis: constant subrules: The predicted vs observed learning curves and distributions of total errors are shown for the subrule groups, (A), C-1, (B) C-2, (C) C-3, and (D) C-4. (See Table 1).

possibility that any simple incremental model will also account for the subrule learning. However, the possibility that the subrule learning is not perfectly all-or-none, that some learning takes place before perfect (criterion) learning, is not ruled out.
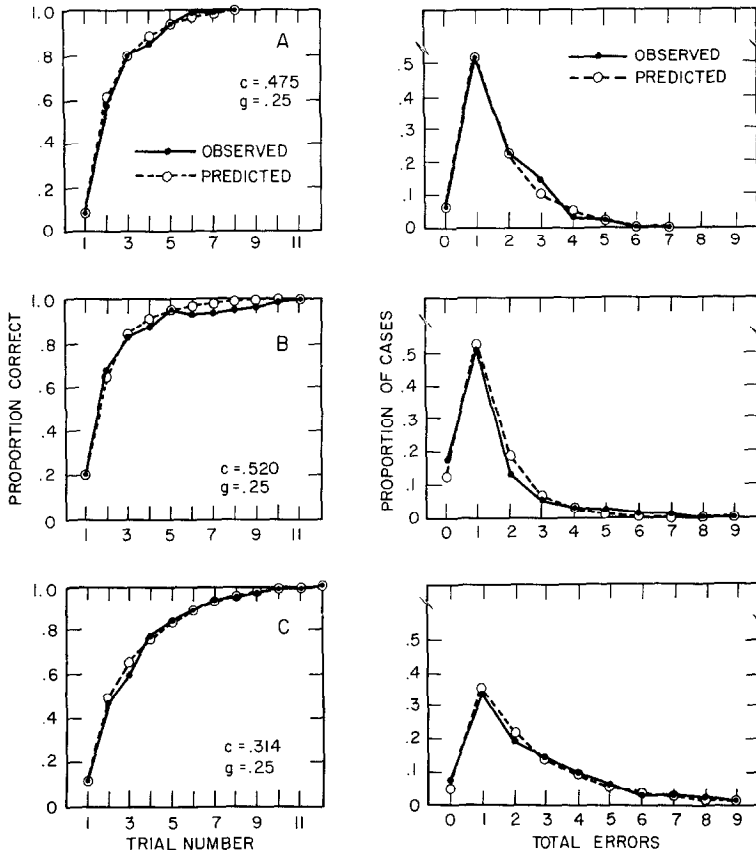


FIG. 4. Subsequence analysis: advancing subrules. The predicted vs observed learning curves and distributions of total errors are shown for the subrule groups, (A) A-1, (B) A-2 (easy), and (C) A-2 (hard). (See Table 1).

## Minimum Chi-Square Analysis: The Insight Model

This section investigates the extent to which there is some learning on the subrules before criterion performance is reached. The development of a model, called the "insight" model (Rumelhart, 1964) provides the framework for this analysis. In its assumptions the insight model lies between the simplest all-or-none model and the

simplest incremental model, and, for certain values of its parameters, reduces to one or the other.

The insight model assumes that upon reinforcement, with probability $c$ the subject learns in an all-or-none fashion, and with probability $1-c$ he learns incrementally.

Mathematically, if $P(C_n)$ is the probability of being correct on the $n$th trial, then with probability $c$,

$$P(C_{n+1}) = P(C_{n+2}) = \cdots = P(C_\alpha) = 1,$$

and with probability $1 - c$,

$$P(C_{n+1}) = P(C_n) + \alpha[1 - P(C_n)].$$

The process starts with $P(C_1) = g$.

When $\alpha = 0$, the insight model reduces to the one-element model, and when $c = 0$, it reduces to the simplest incremental model, the simple linear operator model (Bush and Sternberg, 1959).

The fit of the insight model to the subrule data, grouped according to Table 1, is tested using a minimum chi-square procedure (Atkinson and Crothers, 1964) on trials 2 through 5 of the subsequence protocols. It is reasonable in testing the model to consider only trials 2–5 because (a) the first trial is before the first reinforcement and can be excluded since it reveals nothing of the subrule learning process, and (b) the learning is fast enough on the subrules so that not too much is cut off by stopping with subsequence trial 5. On trials 2–5 any given subject will have one of the 16 possible four-tuples of correct responses and errors: CCCC, CCCE, CCEC, CCEE, CECC, CECE, CEEC, CEEE, ECCC, ECCE, ECEC, ECEE, EECC, EECE, EEEC, EEEE. For any set of values for its parameters, the insight model predicts the probability of each of the 16 error–success four-tuples. If we choose a particular set of values for $\alpha$, $c$, and $g$, the probability distribution across the 16 events can be computed and compared with the observed distribution of the 16 events.

Some information can be gained as to the goodness-of-fit by computing the chi-square value:

$$\chi^2 = \sum \frac{(0 - E)^2}{E},$$

where 0 is the observed proportion of an error–success four-tuple, $E$ is the proportion predicted by the model, and the sum is over the 16 possible four-tuples. The minimum chi-square procedure finds the particular set of parameter values for which this quantity is minimized. In the case of the insight model the procedure does more than just provide the best fit of the insight model to the data. It also indicates through the parameter values yielding the minimum chi-square the degree to which the learning conforms to an all-or-none process. To the extent that $\alpha$ is near zero the learning is all-or-none and to the extent that $c$ is near zero the learning is incremental.

Table 2 presents the results of the minimum chi-square analysis. The table contains, for the insight model, the one-element model, and the simple linear operator model, the minimum chi-square values and minimizing parameter values for each of the seven subrule groups.

TABLE 2

MINIMUM CHI-SQUARE GOODNESS-OF-FIT ANALYSIS[a]

| Subrule group | Insight model | | | | All-or-none model | | | Incremental model | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\chi^2(12df)$ | $\alpha$ | $c$ | $g$ | $\chi^2(13df)$ | $c$ | $g$ | $\chi^2(13df)$ | $\alpha$ | $g$ |
| C-1 | 16.5 | .50 | .18 | .01 | 24.3[b] | .52 | .10 | 17.1 | .23 | .49 |
| C-2 | 37.5[b] | .16 | .43 | .01 | 50.6[b] | .45 | .23 | 199.7[b] | .37 | .23 |
| C-3 | 10.4 | .03 | .67 | .32 | 10.5 | .68 | .34 | 52.9[b] | .60 | .45 |
| C-4 | 3.4 | .05 | .77 | .48 | 3.4 | .78 | .49 | 25.9[b] | .75 | .48 |
| A-1 | 10.6 | .05 | .38 | .26 | 10.8 | .38 | .33 | 48.7[b] | .23 | .49 |
| A-2E | 20.7 | .00 | .50 | .20 | 20.7 | .50 | .20 | 200.1[b] | .20 | .56 |
| A-2H | 12.3 | .00 | .50 | .20 | 12.4 | .32 | .21 | 209.5[b] | .20 | .38 |

[a] The chi-square values reflect the deviation between predicted and observed frequencies of the 16 possible error–success four-tuples on trials 2–5.
[b] Significant.

For five of the seven subrule groups in Table 2 the minimizing value of $\alpha$ in the insight model is near zero; there appears to be negligible learning of these subrules before perfect learning is attained. This lack of partial learning is emphasized in that, when $\alpha$ is set to zero, the resulting two-parameter one-element model fits the four-tuple data as well as the three-parameter insight model. The data for groups C-1 and C-2, which are not fit well by the all-or-none model, are atypical for the experimental reasons given earlier. It would be very difficult statistically to assert that the subrule learning is anything more complicated than all-or-none.

DISCUSSION

The results of the preceding analysis of sequence learning are summarized by three findings. (a) Overall, the learning of the experimental sequences appears very complex. (b) However, when the learning protocols are decomposed according to the subrules underlying the sequences, the resultant subrule learning curves are very orderly. (c) And finally, an all-or-none model of the subrule learning accounts very well for the obtained subrule data.

The conceptual framework implicit to the model analysis of the results views sequence learning as an association task. In order to learn the constituent subrules of a sequence, and hence the sequence, subjects must associate arithmetic operations with cues for when to apply the operations. The stimulus and response units of the association, however, are more complex than those which characterize conditioning or simple paired-associate learning; both the arithmetic operations and the cues for applying the operations must be abstracted from the progressing sequence.

In terms of this associative conception of sequence learning the results of the data analysis have two main implications which merit explicit comment.

1. The breakdown of the sequence learning into component orderly subprocesses illustrates that apparently complex performance can result from combinations of simple learning processes. In particular, if learning a subrule requires a relatively straightforward association, complex performance can reflect the learning of several concurrent associations.

2. The all-or-none nature of the subrule learning implies that, in spite of the more complex nature of the stimulus and response involved in learning a subrule, the association is formed in a fashion which is formally similar to conditioning or simple verbal association.

It would be an unwarranted extrapolation from this study to assume that any sequence learning task with a subrule structure should exhibit all-or-none learning of the subrules. When paired-associate experiments are complicated along one or more of several dimensions, e.g., amount of immediate memory or response integration required, the learning becomes progressively less likely to conform to an all-or-none model. Thus, sequences which demand more of the subject's memory for past members or more response integration might be expected to result in a departure from all-or-none learning.

## REFERENCES

ATKINSON, R. C., AND CROTHERS, E. J. A comparison of paired-associate learning models having different learning and retention axioms. *Journal of Mathematical Psychology*, 1964, 1, 285–315.

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

BOWER, G. H., AND THEIOS, J. A learning model for discrete performance levels. In R. C. Atkinson (Ed.), *Studies in mathematical psychology*. Stanford: Stanford Univer. Press, 1964. Pp. 1–31.

BUSH, R. R., AND STERNBERG, S. H. A single-operator model. In R. R. Bush and W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univer. Press, 1959. Pp. 204–214.

ESTES, W. K. All-or-none processes in learning and retention. *American Psychologist*, 1964, **19**, 16–25.

HOLLAND, P. W. A variation on the minimum chi-square test. Tech. Rep. No. 35, Institute for Mathematical Studies in the Social Sciences, Stanford University, 1965.

RESTLE, F. Sources of difficulty in learning paired associates. In R. C. Atkinson (Ed.), *Studies in mathematical psychology*. Stanford: Stanford Univer. Press, 1964, Pp. 116–172.

RUMELHART, D. E. An insight model for learning. Unpublished paper, 1964.