2

Concept identification

GORDON H. BOWER and THOMAS R. TRABASSO, *Stanford University*

The learning and use of categories (concepts) represents an elementary and general form of cognition by which man adjusts to his environment. As a conceptualizing, classifying animal, man uses language, and the basic units of his language and thoughts are concepts. Although the logical and epistemological study of concepts and their usages is quite advanced, the psychological analysis of conceptual behavior is in a rudimentary stage. Thus a challenging task for present-day experimental psychologists is to construct useful analyses and explanations of how people learn to categorize and classify their social and physical environment.

The development of concepts in young children has been studied by a number of psychologists in attempts to describe the reinforcing environment that influences a child's use of class names such as bird, boy, and book in appropriate fashion. In schematic outline, at least, we are reasonably certain of the variables that produce and control such learning. At the other extreme, our knowledge is less than adequate in explaining how certain adults develop or form (some would say "create") concepts that are novel and useful representations of nature (e.g., the concept of the gene).

Much is known about how to teach concepts to others. In the past decade, a large number of experimental facts have come to light concerning variables that influence and control the speed with which an adult subject learns a particular concept. It is our contention, however, that with very few exceptions, these facts have been derived from explorations of the process of *concept selection* rather than of *concept formation*. Concept formation refers to the initial development or learning of an equivalence class *de novo*, whereas concept identification refers to the selection of the appropriate classification from among a set of attributes already known by the learner. This distinction

This research was supported by a research grant, M-3849, to the first author from the National Institutes of Mental Health. The same agency provided the second author with a postdoctoral fellowship, MPD-18070.

Reprinted from STUDIES IN MATHEMATICAL PSYCHOLOGY, edited by R. C. Atkinson. Stanford University Press.

has been proposed and amplified by Hunt (1962). Most psychologists study how quickly a subject comes to classify his laboratory environment in the same way that the experimenter has chosen to classify it. Thus, the subject is not usually acquiring novel concepts. In most instances, he already has in his repertoire the perceptual discriminations and names of attributes that are relevant to the solution of the problem. His main job is one of selection, through a trial and error process, of the appropriate classifications that the experimenter has decided to reinforce.

This paper presents a theoretical analysis of concept-identification experiments and reports some new experimental results pertinent to the analysis. We deal exclusively with what has been called two-category concept identification, in which the stimulus patterns are constructed from two-valued dimensions or attributes and the subject makes a binary classification of each pattern by using one of two mutually exclusive responses. The logical structure of such problems can best be seen by reference to Table 1, which presents, in a schematic way, the method for construction of the 2^N patterns from N independent binary dimensions. Table 1 illustrates a population of stimuli constructed from three stimulus dimensions. For concreteness, we may suppose that the three stimulus dimensions are the color (red or blue), shape (circle or square), and size (large or small) of geometric patterns. The two values of each dimension are represented by the entries 1 and 0 under each stimulus column.

TABLE 1
SCHEMATIC OUTLINE OF STIMULUS CONSTRUCTION FOR CONCEPT IDENTIFICATION

D1	D2	D3	Response Assignment
1	1	1	A
1	1	0	A
1	0	1	A
1	0	0	A
0	1	1	B
0	1	0	B
0	0	1	B
0	0	0	B

Each row of Table 1 represents one of the 2^3 possible stimulus patterns that would be presented to a subject learning this concept-identification problem. The responses A and B are assigned according to a systematic rule: if the value of D1 is 1, the answer is A; if the value of D1 is 0, the answer is B. In this example, D1 is called the *relevant* stimulus dimension since the classificatory response is perfectly correlated with D1 values. Dimensions D2 and D3 are independent *irrelevant* dimensions since their values appear

equally often as A's and as B's. In an experimental realization of this problem, each pattern is presented, one at a time, and the subject attempts to classify it as an A or B; following his response, the experimenter provides information about the correct classification. The series of 2^N patterns is gone through repeatedly in random order until the subject achieves a criterion of learning, such as 15 consecutive correct responses.

The systematic rule in Table 1 is but one of many rules that could be used for assigning the responses A and B to the eight patterns. A recent monograph by Shepard, Hovland, and Jenkins (1961) has explored six different types of response assignments, which are ordered in their difficulty of being learned. The symmetric assignment in Table 1 was called a "Type I" classification, and it was the easiest for subjects to learn. All of our studies are of Type I concepts; later we shall discuss the other types of response assignments studied by Shepard *et al.* (1961).

In carrying out such experiments, we usually employ stimulus dimensions that are perceived as obvious attributes by nearly all subjects in our population. Thus, the basic perceptual differentiations are assumed to be already part of the subject's repertoire when he enters the experiment. Moreover, the values within a particular dimension are usually clearly discriminable. From the subject's point of view, the presence of a given stimulus dimension depends, in fact, upon there being distinguishable values of it appearing over the series; otherwise, the values of that dimension have no cue properties and in no way will influence the subject's behavior.

A variety of theories has been proposed to explain and account for the major phenomena of two-category concept-identification experiments. These theories may be broadly classified as incremental or all-or-none, according to their expectations regarding the systematic changes that take place in the subject's behavior during the course of the experiment. Theories of the incremental sort are traditional and more plentiful; the all-or-none theories are of more recent vintage. Probably the clearest recent statement of an incremental theory is that provided in an article by Bourne and Restle (1959). They explicitly view concept-identification experiments as simply another class of discrimination studies, and proceed to apply Restle's (1955) theory of discrimination learning to such data. This theory postulates two hypothetical processes: the conditioning of values of the relevant cues to the corresponding correct responses, and the gradual adaptation or elimination of irrelevant cues as determinants of response tendencies. Both processes are conceived to operate trial by trial at a fixed rate, θ. In a two-response problem, the probability that one of the R relevant cues, k, is conditioned to its correct response by trial n will be

$$C(k, n) = 1 - \tfrac{1}{2}(1 - \theta)^{n-1},$$

and the probability that one of the I irrelevant cues is adapted by trial n is

$$A(k, n) = 1 - (1 - \theta)^{n-1}.$$

The probability of a correct response on trial n is given by the average probability that an unadapted cue is conditioned, or

$$P_n = \sum_k \frac{[1 - A(k, n)]C(k, n)}{\sum_k [1 - A(k, n)]}. \tag{1}$$

After substitution and some simplification, Eq. (1) can be reduced to the following final form:

$$P_n = 1 - \frac{\frac{1}{2}(1 - \theta)^{n-1}}{r + (1 - r)(1 - \theta)^{n-1}}, \tag{2}$$

where r is the proportion of relevant cues in the problem. Restle further adds the assumption that $r = \theta$. This assumption leaves only one free parameter and has the theoretical advantage of directly relating the learning rate, θ, to a structural aspect of the stimuli composing the problem. Using this theory, Bourne and Restle (1959) were able to bring a remarkable degree of orderliness into a wide range of data.

The Bourne-Restle model incorporates an incremental theory because it is assumed that each subject changes his response probability in accordance with the values specified by Eq. (2). If θ is sufficiently small, as it usually is, then each individual's response probability will change over a large number of values between the initial value of $\frac{1}{2}$ and the asymptotic value of unity. The sole source of variance for the model is in the customary binomial variance of sampling associated with a given value of P_n.

The alternative theories we shall discuss in this paper have as their basic postulate the notion that the subject's performance changes in discrete, discontinuous steps; moreover, it is assumed that the performance of an individual subject can be characterized by assigning to him one of two possible values of response probability: an initial value, p, usually near the chance level, and a terminal value of unity. The appropriate mathematical model for such theories is a two-state Markov chain. The states refer to the possible values of response probability, p or 1. The subject starts in the initial state, where he has probability p of a correct response; as information accrues to him, some event occurs that changes his probability of a correct response to 1.00. For the moment, we need not specify this theory any further, since it is possible to distinguish it empirically from incremental theories. After we have shown this empirical distinction and have given some evidence pertinent to deciding the issue, we return to filling out the psychological rationale behind the all-or-none models.

The critical distinction between the all-or-none and incremental theories is simply this: according to the all-or-none theory, the performance of an individual subject should show no improvement over trials before he learns; according to an incremental theory, the performance of the individual subject should improve monotonically with practice. The issue may be decided by inspection of those trials before the last error of a given subject. According

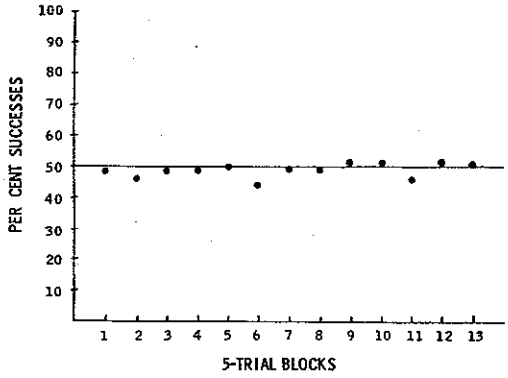


FIG. 1. Stationarity data: percentage of successes prior to the last error plotted in blocks of five trials (Data from Trabasso, 1961).

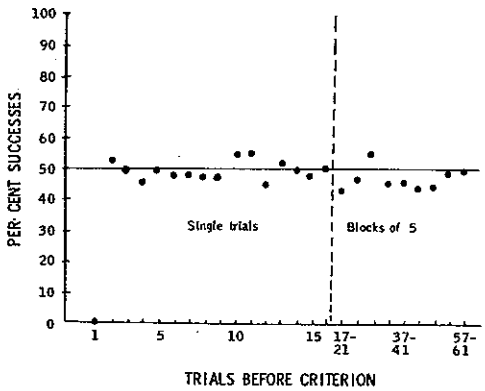


FIG. 2. Backwards learning curve: percentage of successes plotted over trials before the last error (Data from Trabasso, 1961).

to the all-or-none theory, the probability of a correct response over trials before the last error should remain approximately constant near the chance level; according to the incremental theory, it should increase monotonically. Suppes and Ginsberg (1963) noted this critical distinction between all-or-none and incremental theories, and they inferred from stationary presolution curves that children's learning of certain mathematical concepts could be described by an all-or-none model.

We will present briefly two sets of experimental results on trials prior to the final error; other sets of results showing the same feature will be scattered throughout later parts of the paper. The first set of results is taken from Trabasso's (1961) thesis experiment. Over 200 subjects were trained on a number of two-choice single classification problems. The stimulus materials consisted of flower designs (taken from Hovland, 1953) that varied in a number of attributes, such as the type of flower, its color, number and serration of leaves, etc. These flower designs were to be classified as one of two types. Different subjects had differing combinations of relevant attributes; this affects the learning rate on a problem but is immaterial to deciding whether performance is constant prior to a subject's final error. Hence, we

have pooled all subjects for making our analysis. The relevant results are shown in Fig. 1, giving the proportion of correct responses in five-trial blocks; each point plotted involves only those subjects who made their final error on some later block. As subjects make their final error (and learn), they are dropped from consideration; hence the number of subjects involved decreases over successive trials. The number of observations is quite large, averaging around 550 responses per block.

The most striking feature of the successive estimates plotted in Fig. 1 is their constancy over the 65 trials included in the graph. A plot of the backwards learning curve is shown in Fig. 2. This curve contains only subjects who solved the problem (gave 10 consecutive correct responses). The number of observations is quite large (e.g., 213 and 184 on trials 2 and 3, respectively), and again the points are stationary. The probability of a correct response on the two trials preceding the last error is .53 and .50, respectively, neither of which differs from the a priori chance level of .50.

A statistical test for the constancy of the estimates in Fig. 1 has been proposed by Suppes and Ginsberg (1961), and involves calculation of a χ^2 value. For k prelearning trial blocks, the pooled data are arrayed in a $2 \times k$ table (errors and successes by trials). The number of observations in each trial block decreases over blocks. The expected probability of an error is obtained from the ratio of total errors to the total observations in the $2 \times k$ table. The expected frequency for each cell is calculated by multiplying the observations in that trial block by the over-all mean percentage of errors or successes as the case may be. Then the χ^2 value is calculated in the conventional manner from this $2 \times k$ table of observed and expected frequencies. In the present case for Fig. 1, the test yields a χ^2 of 14.41, which, with 12 degrees of freedom, gives a $P > .20$. Thus, there is no evidence for the improvement in performance (before the last error) that is predicted by incremental theories. Instead the picture that emerges is that response probability is constant for a number of trials before the subject learns on a single trial.

One other piece of evidence from Trabasso's data will be mentioned since it is of critical importance to the all-or-none theories. The evidence concerns the independence of successive responses prior to the last error. According to the model sketched above, we expect the sequence of responses the subject generates before he learns to be an independent Bernoulli series with parameter p of a correct response. The independence assumption means that the probability of a correct response on trial $n + 1$ should be the same whether a correct or incorrect classification occurred on trial n. Suppes and Ginsberg (1961) have proposed another χ^2 test for this independence property of the data. One counts the number of times a success or failure is followed on the next trial by a success or failure, summing frequencies over all trials before a subject's last error and summing over all subjects. The 2×2 table for the data represented in Fig. 1 is shown below in Table 2. The conditional probabilities of a success following a success or failure on trial n are about equal;

TABLE 2
TRANSITION FREQUENCIES FOR INDEPENDENCE TEST: FREQUENCY OF SUCCESS OR FAILURE ON TRIAL $n + 1$ GIVEN A SUCCESS OR FAILURE ON TRIAL n
(Data from Trabasso, 1961)

		Trial $n + 1$ Success	Trial $n + 1$ Failure	Conditional Probability of Success on $n + 1$
Trial n	Success	1575	1594	.497
	Failure	1614	1697	.487

the χ^2 is .59, which, with one degree of freedom, gives a $P > .30$. This test has considerable power, being based on a total of 6,480 observations. Thus, one cannot reject the assumption that successive responses prior to the last error are statistically independent.

A second illustration of the stationarity predicted by the all-or-none theory is taken from an experiment by Bower (1962b). Twenty-five subjects learned to a criterion of 15 consecutive correct responses the two-choice problem illustrated in Table 3. The stimuli were five-letter consonant clusters. One of two letters appeared in each of the five positions in the left-to-right order. An example is shown in the middle of Table 3. There are 2^5 or 32 different consonant clusters that can be constructed from this array. The solution to the problem depended on the letter in the fourth position: if it was *R*, the answer was 1; if it was *Q*, the answer was 2.

The results relevant to the stationarity assumption are given in Fig. 3, which shows the proportion of correct responses in blocks of four trials prior to the last error for any given subject. Figure 3 was constructed in the same manner as was Fig. 1. Again the successive estimates show no trend away from the initial level. The test for stationarity yielded a χ^2 of 11.80 (with 14 df), which gives $P > .50$. Analysis of conditional probabilities, to test the independence assumption, gave results comparable to those reported in

TABLE 3
STIMULUS MATERIALS IN THE BOWER (1962b) STUDY

	Positions 1	2	3	4	5	
Consonant Letter Pairs	(J, H)	(X, V)	(T, K)	(R, Q)	(L, Z)	
Example:	J	V	K	R	Z	
Solution:	—	—	—	R	—	Response 1
	—	—	—	Q	—	Response 2

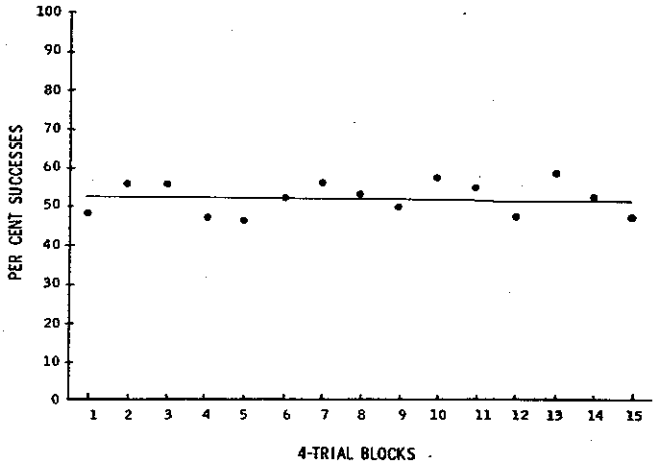


FIG. 3. Stationarity data: percentage of successes prior to the last error plotted in blocks of four trials (Data from Bower, 1962).

Table 2 for the Trabasso data. Again one may conclude that the responses prior to the final error can be represented by a stationary series of independent Bernoulli observations.

1. Psychological interpretations of all-or-none models

We have seen in the preceding section that a subject's performance with respect to the correct classification can be described quantitatively as a two-state process, corresponding to response probabilities of p and 1.00. Accordingly, the psychological interpretations to be considered are restricted principally to those that imply this two-state process at the level of response probabilities.

One theory meeting this requirement has been proposed recently by Restle (1962). He calls it the "strategy-selection" theory. The subject in a concept-identification experiment is viewed as testing out various hypotheses (strategies) about the solution to the problem. Each problem defines for the subject a population of hypotheses. The subject samples one of these hypotheses at random and makes the response dictated by the hypothesis. If his response is correct, he continues to use that hypothesis for the next trial; if his response is incorrect, then he resamples (with replacement) from the pool of hypotheses. The hypotheses are partitioned into a proportion c of correct hypotheses that always lead to a correct response, a proportion w of wrong hypotheses that always lead to an error, and the remaining proportion i of irrelevant hypotheses that lead to correct or incorrect responses randomly with probabilities of one-half. Restle has related the proportion of correct strategies, c, to the proportion of relevant cues and other structural aspects

of the stimuli defining the population; these hypotheses will be discussed later.

Restle's theory leads to an all-or-none model at the level of response probabilities. Prior to the trial on which the subject selects a correct hypothesis (by definition, his last error trial), there is no improvement in his average probability of a correct classification. The probability distribution of the total number of errors a subject makes before selecting a correct strategy and reaching criterion is derived easily. The probability that any given error is the last one is c, the likelihood that upon resampling the subject selects a correct hypothesis. The probability that the first hypothesis selected will be correct so that no errors are made is also c. Hence, the probability of exactly k errors is $c(1-c)^k$, the geometric distribution, with mean $(1-c)/c$ and variance $(1-c)/c^2$. An error is an uncertain *recurrent event* in this theory; each error resets the process back to zero, where it starts over again from scratch. This aspect of the theory follows from the assumption that the subject samples hypotheses randomly and with no memory of what previous hypotheses have been tried and rejected (sampling with replacement). If hypotheses were being cast out of the population as they were sampled, tested, and rejected, and if no new hypotheses were added to the initial pool, then the proportion of correct strategies, c, would increase with each error. In this case, the probability of selecting a correct strategy after the nth error would be

$$c_n = \frac{C}{C + W + I - n} \qquad (0 \leq n \leq I + W)$$

$$= \frac{c_0}{1 - n/H},$$

where C, W, and I are the initial number of correct, wrong, and irrelevant strategies, and $C + I + W = H$. The probability that the nth error is the last one increases with n. The probability of exactly k errors is

$$\Pr(T = k) = \frac{c_0}{1 - k/H} \prod_{j=0}^{k-1} \left(1 - \frac{c_0}{1 - j/H}\right).$$

If H is assumed to be large, then sampling with or without replacement makes little material difference in the outcomes expected by the two assumptions. No available data are sufficiently free of sampling error to permit quantitative decisions between the constant-c assumption and the increasing-c assumption with large H. However, some of the results reported later in this paper contradict the notion that c changes appreciably over trials.

The psychological interpretation we propose is but a slight variation on Restle's hypothesis-testing model. We will interpret our results primarily in terms of perceptual acts (attention) and conditioning of cues to which the subject attends. Thus, we will speak of a stimulus-selection process whereby the subject comes to attend principally to the values of the relevant stimulus dimension, and of a conditioning process whereby he learns, in paired-

associate fashion, the associations between the various values of the relevant dimension and their assigned responses. In the usual two-choice concept-identification experiment, the latter, paired-associate phase, is apparently very rapid, probably occurring in a single trial. With binary relevant dimensions and a Type I classification, it is really necessary for the subject to form only a single association, between one value and its response, in order to start a criterion run to correct responses. If, say, color (red or blue) is the relevant dimension and the subject learns that red things are A's, by the symmetry of the situation he will respond with B to non-red (blue) things. During the criterion run of correct responses, he can, of course, learn the blue-B association as well. Thus, in applying the theory to two-choice problems, it will be assumed that the stimulus-selection process is being observed throughout, with the necessary paired-associate learning occurring very quickly at the end.

It is clear that the relative difficulty of the stimulus-selection and paired-associate phases can be manipulated experimentally. To lengthen the paired-associate phase, one increases the number of specific values in the relevant dimension, each to be associated with a unique identifying response. Correspondingly, the stimulus-selection phase can be shortened by reducing the number of irrelevant attributes. The limit of this process is a paired-associate task, in which there are no irrelevant cues and the subject's sole task is to form associations between the values of the relevant attribute and the assigned responses. The theory thus identifies an experimental continuum from concept identification to paired-associate learning. A generalization of the model to handle experiments along this continuum is given in Appendix A. No research has yet been done along intermediate points on this experimental continuum.[1] It might be mentioned incidentally that the theoretical unit of analysis changes as one proceeds along this continuum. In paired associates, the unit of analysis is the sequence of responses to the individual stimulus item. In our analysis of concept identification, the unit to which the theory applies is the sequence of responses to the entire series of patterns, without distinguishing whether the pattern on trial n contained the value 1 or 0 of the relevant dimension. From a subject learning a set of 2^N distinct paired-associate items, one obtains 2^N response sequences for analysis; by contrast, for a subject learning a concept-identification problem with 2^N overlapping patterns, one obtains only one sequence for theoretical analysis. This difference in the unit of analysis accounts in large part for the quantitative differences in "total errors," "learning rates," etc., when the same model is applied to the two experiments.

---

[1] Shepard, Hovland, and Jenkins (1961) have suggested another continuum between concept identification and paired-associates which depends upon the assignment of two responses to the 2^N stimulus patterns. The continuum we identify here is one over which the number of irrelevant attributes, number of values of each attribute, and number of responses all vary, but there is a one-to-one assignment of the responses to the values of the relevant attribute. It would be a generalized Type I problem, in the terminology of Shepard *et al.*

To return to the discussion of two-category problems: We have assumed that we are observing primarily a stimulus-selection process in the subject's behavior. The stimulus selection is conceived to proceed randomly, according to certain structural parameters, until the subject both selects *and* conditions a relevant cue to its appropriate response. Once this joint event occurs, the subject henceforth uses the cue as a basis for responding and the problem is solved. Prior to the occurrence of this critical event, the subject responds systematically to unconditioned relevant cues or to conditioned or unconditioned irrelevant cues, being correct with some chance probability p and in error with probability $q = 1 - p$.

Each stimulus attribute i is represented by a measure or weight w_i. These weights may be interpreted as the attention value of attribute i or as summarizing all factors determining the subject's selection of attribute i as a stimulus to be tested. We will let w_r represent the weight of the relevant attribute and let w_i represent the combined weight of the irrelevant attributes. The relative weight of the relevant attribute defines a structural parameter, r, as

$$r = \frac{w_r}{w_r + w_i}, \tag{3}$$

which characterizes a particular problem. The parameter r is also the probability that, in a random search of attributes, the subject will focus his attention on the relevant attribute; r also may be described as the probability of sampling the relevant cue in a one-element sample. Given that the subject selects a relevant attribute, there is some probability, θ, that he conditions its value to the reinforced response. Thus, $r \cdot \theta$ is the probability that the critical learning event happens on any particular trial; in an all-or-none theory, $r\theta$ is the probability that on any particular trial the subject leaves the initial state and goes into the terminal state, where he gives errorless performance. In other contexts we would identify $c = r\theta$ with the net "learning rate." We here identify c as the product of two manipulable parameters: r, which reflects aspects of the subject's perceptual processing of the stimulus materials and varies with the weight of the relevant and irrelevant cues, the discriminability of the two values of the relevant dimension, prior instructions, pretraining of the subjects, and so forth; and θ, which is the conditioning parameter governing association learning of values to responses, which will vary with reinforcement variables such as completeness and immediacy of feedback information. Explicit tests of Eq. (3) will be given below, in the section on additivity of cues.

From an abstract point of view, the model we have outlined has the following formal properties: the subject begins in some initial state—call it U for unlearned—where he has probability p of a correct response. On each reinforced trial there is some probability c that he moves into a terminal, absorbing state C (conditioned or correct) where his probability of a correct response is 1.00. Stated in this way, the model is identical to the one Bower

(1961) has used for analysis of elementary forms of paired-associate learning, and its mathematical properties are known in detail.

This model was used to predict details of the data from the 25 subjects in the previously discussed experiment that involved five-letter consonant clusters as stimuli. The estimate of p was the observed proportion of successes prior to the last error, which was .523. The estimate of the learning rate, c, was obtained from the mean errors to criterion. The following rationale for the estimate may help the reader: If the subject has probability c of learning on each trial, then on the average there will be $1/c$ trials before the effective learning event occurs; for each of these $1/c$ prelearning trials, there is a probability $q = 1 - p$ that the subject will make an error; hence, his expected number of errors is q/c. Equating $.477/c$ to the observed mean errors of 12.16, we obtain the estimate $c = .039$. Referring then to the formulas given in Bower's paper (1961), we obtain predictions and compare them with observed values (Table 4).

The predictions fit the detailed data fairly well. It will be noticed that the model predicts and one observes large variance in most statistics. With such variability and with a small sample of 25 subjects, the predictions are as close as can be expected. None of the predictions differ significantly from corresponding observed values by t or F tests of statistical significance. Results of this nature provide encouraging signs for mathematical models of concept identification. In effect, the results establish the depth of analysis and the

TABLE 4
CONSONANT-CLUSTER EXPERIMENT: DETAILED PREDICTIONS

Statistic	Observed	Predicted
Average errors	12.16	12.16
Standard deviation	12.22	12.18
Errors before first success	.92	.89
Standard deviation	.98	1.14
Average trial of last error	25.70	24.50
Standard deviation	28.90	25.00
Probability of an error following an error	.47	.46
Runs of errors	6.44	6.57
Runs of 1 error	3.62	3.57
2 errors	1.32	1.63
3 errors	0.64	0.75
4 errors	0.40	0.35
Alternations of success and failure	12.33	12.41
Error-error pairs		
1 trial apart	5.76	5.45
2 trials apart	5.04	5.22

degree of predictive accuracy that one could routinely expect for further tests of models for concept identification.

2. Additivity of cues

A variety of experiments on concept identification have yielded a body of systematic results relating the speed of learning to the stimulus structure of the problem. The purpose of this section is to show how the theory makes contact with this body of facts.

If the successful solution of a concept problem requires the selection of the relevant cue, the probability of this selection will vary directly with the numbers and weights of the relevant cues and will vary inversely with the numbers and weights of the irrelevant cues. This relation is formalized in the assumption about the net learning rate:

$$c = \frac{\theta \sum_{i \in R} w_i}{\sum_j w_j}, \tag{4}$$

where the sum in the numerator is over the set of R redundant relevant dimensions and the sum in the denominator is over all cues in the stimulus pattern. Equation (4) is adopted directly from Restle's earlier theoretical work (1955, 1957). We shall show how Eq. (4) permits some free predictions in several different "cue-additivity" paradigms, and then we shall test these predictions by some experimental results of Trabasso (1962) and Bourne and Haygood (1959). In all cases, the estimated or predicted c's are related to mean errors by the expression $E(T) = q/c$, where we assume that $q = \frac{1}{2}$ for the two-choice problems.

The first type of additivity experiments we shall consider employ a constant pool of stimulus dimensions 1, 2, 3, $\cdots$, N that are always present for all subjects. In condition 1, subjects learn with cue 1 relevant and the remainder irrelevant; in condition 2, subjects learn with cue 2 relevant; and in condition 3, subjects learn with both cues 1 and 2 redundant and relevant. The theory relates the learning rate in condition 3, call it c_3, to the learning rates, c_1 and c_2, estimated from conditions 1 and 2. The relation is $c_3 = c_1 + c_2$.

PROOF. $$c_1 = \theta \frac{w_1}{w_1 + w_2 + w_i}, \qquad c_2 = \theta \frac{w_2}{w_1 + w_2 + w_i},$$

$$c_3 = \frac{(w_1 + w_2)}{\theta w_1 + w_2 + w_i} = c_1 + c_2, \tag{5}$$

where w_i is the combined weights of all other cues in the pool. By elementary algebra the implied relation between average total error scores is

$$T_3 = \frac{T_1 T_2}{T_1 + T_2}. \tag{6}$$

PROOF. $$c_1 = \frac{q}{T_1}, \qquad c_2 = \frac{q}{T_2}$$

$$T_3 = \frac{q}{c_3} = \frac{q}{c_1 + c_2} = \frac{q}{q/T_1 + q/T_2} = \frac{T_1 T_2}{T_1 + T_2}.$$

Adequate tests of Eqs. (5) and (6) are not available in the concept-identification literature known to the writers. However, rather than present no results at all, we will illustrate the formula by predicting some results on the discrimination-learning of animals. The material is intended to be purely illustrative and is not advanced to support the idea that rats and humans go about learning discrimination problems in the same way. The original plan for using such data for testing additivity predictions was published in an article by Restle (1957).

The first set of results we shall discuss were reported by Scharlock (1955). He trained rats in a cross-maze having the end of one arm illuminated by a light bulb and the other arm darkened. Three conditions will be considered: "place-learning" rats that were started alternately at the north and south ends of the cross maze and were rewarded for choosing the lighted arm; "response-learning" rats that had alternate starting places and were rewarded for turning right; and "place-plus-response-learning" rats that were always started from the north end of the maze and were rewarded for going to the lighted arm, which always involved a right-turning response. Each condition contained 20 subjects. Average total errors before learning for the three conditions were 9.7 for place learners, 6.7 for response learners, and 4.0 for place-plus-response learners. Substitution of 9.7 and 6.7 for T_1 and T_2 into Eq. (6) leads to the prediction of 3.97 errors for the place-plus-response learners. This prediction is close to the observed value of 4.0 errors. A second example of the prediction of Eq. (6) uses data reported by Warren (1959). Cats were trained on discrimination problems in which position, object, or position-plus-object were relevant cues. Observed mean error was 11.13 on the position problem and 28.63 on the object discrimination. Errors on the problem with both position and objects as relevant cues averaged 8.00, whereas Eq. (6) predicts a value of 8.03.

The second type of cue-additivity experiments we shall consider are those in which the stimulus situation is modified by the addition of new relevant or irrelevant cues. We will consider two experiments of this type. The first, reported by Bourne and Haygood (1959), is based on the assumption that all cues in their experiment had equal weights; the second, reported by Trabasso (1961), uses the theory to estimate the weights of the added cues.

Bourne and Haygood trained groups of 10 subjects in each of 12 different conditions of a two-category concept-identification problem. Geometric forms constituted the stimuli; other cues that could be present were color, size, number, borders, etc. The 12 conditions of the experiment are listed in Table 5, along with the average errors made by the 10 subjects who learned a

given problem. The values under columns R and I give the number of relevant and irrelevant attributes for subjects working on a given problem.

Bourne and Haygood interchanged from one subject to the next the particular cues (relevant and irrelevant) associated with a given problem. By this balancing procedure they hoped to average out any unique effects due to the use of particular cues in the various conditions. From the reported mean errors of subjects in a given condition, one cannot estimate the relative weights of the variety of cues involved. In order to generate predictions for these data, we have made the simplifying assumption that all attributes have equal weight. With this assumption, the learning rate for a problem with R relevant and I irrelevant cues is

$$c = \frac{\theta R}{R + I}. \tag{7}$$

The expression for average total errors is

$$T = \frac{.5}{c} = \frac{.5}{\theta}\frac{(R + I)}{R} = K\frac{(R + I)}{R}. \tag{8}$$

Using the observed values of T given in Table 5, we obtained a least-squares estimate of $K = 2.24$. Equation (8) was then used to generate the all-or-none predictions listed in Table 5. We have listed also the predictions of these numbers from Restle's incremental theory, which were published in the Bourne-Restle paper (1959). The number of arbitrary parameters estimated for the Bourne-Restle predictions was three; for the all-or-none theory, one

TABLE 5

ADDITIVITY OF CUES PREDICTIONS

(Data from Bourne and Haygood, 1959; incremental prediction from Bourne and Restle, 1959)

R	I	T obser.	Predicted T	
			All-or-none	Incremental
1	1	4.3	4.48	5.5
2	1	3.2	3.37	3.9
3	1	3.1	3.00	3.2
4	1	3.1	2.81	2.8
5	1	2.7	2.69	2.7
6	1	2.1	2.61	2.5
1	3	8.2	8.97	8.2
2	3	6.5	5.61	5.7
3	3	5.7	4.48	4.4
4	3	3.9	3.93	3.8
1	5	13.6	13.47	11.8
2	5	7.4	7.85	7.8

parameter (K) was estimated. The sum of squared deviations, observed minus predicted errors, is 3.51 for the all-or-none model, and is over twice as large (7.93) for the incremental model.

We now use the data from Trabasso's experiment (1961) to illustrate the use of the model to estimate the weights of added cues. The seven experimental conditions of immediate interest are listed in Table 6. Each condition contained 20 subjects. The stimuli were flower designs that varied in a number of attributes such as the angle of the leaves to the stem, the color of the flower, the color of the angle, the type of flower, number and serrations of leaves, etc. All subjects had the same set of irrelevant cues; to this common pool were added the relevant attributes listed in Table 6 for the various experimental conditions. We will not describe the experimental procedure in any more detail here since further details may be obtained by referring to Trabasso's report (1963). In Table 6 we have also listed for each condition the mean errors, T, the number of subjects solving the problem within 65 trials, N_s, the estimate of c based on T and N_s, and, where applicable, the predicted c and T for the additive groups.

Conditions 5 and 7 involve the same type of additivity procedure, and we derive the relation for the general case. Throughout these derivations with the all-or-none model, θ is assumed to equal 1.00; this relieves us of a degree of freedom but does permit predictions to be made. We let w_i represent the combined weight of the common pool of irrelevant cues and let w_a and w_b represent the weights of the added relevant cues. The relations are as follows:

(9a) $$c_a = \frac{w_a}{w_a + w_i} \qquad \text{with } a \text{ added,}$$

(9b) $$c_b = \frac{w_b}{w_b + w_i} \qquad \text{with } b \text{ added,}$$

(9c) $$c_{ab} = \frac{w_a + w_b}{w_a + w_b + w_i} \qquad \text{with } a \text{ and } b \text{ added.}$$

Equations (9a) and (9b) can be solved for w_a and w_b in terms of w_i and the c's. Substitution for these in Eq. (9c) leads to the following relation:

(10) $$c_{ab} = \frac{c_a + c_b - 2c_ac_b}{1 - c_ac_b}.$$

Equation (10) was used in obtaining predictions for conditions 5 and 7 in Table 6. Inspection shows the predictions to be reasonably close.

The prediction for condition 6 differs slightly because one of the added cues (say, cue b) is irrelevant. For this reason, Eq. (9c) is replaced as follows:

(11a) $$c_{a/b} = \frac{w_a}{w_a + w_b + w_i}.$$

Substitution for w_a and w_b in terms of Eqs. (9a) and (9b) leads to the result

$$c_{a/b} = \frac{c_a(1 - c_b)}{1 - c_a c_b}. \tag{11b}$$

Equation (11b) was used to predict c and mean errors for subjects learning under condition 6. The observed c estimate was .029, while the predicted c was .032, not significantly different.

TABLE 6
ADDITIVITY OF CUES PREDICTIONS
(Observed Data from Trabasso, 1961)

Experimental Condition	Obs. $E(T)$	N_s	Est. c	Pred. c	Pred. $E(T)$
1. Angle cue relevant	19.50	14	.0181		
2. Angle, emphasized red	12.45	18	.0363		
3. Flower-color rel.	3.40	19	.1407		
4. Angle-color rel.	4.05	19	.1180		
5. Angle + ang-color rel.	3.45	20	.1450	.145[a]	3.45[a]
6. Angle, with angle-color irrel.	14.65	17	.0292	.032[a]	15.50[a]
7. Angle + flower-color rel.	2.40	20	.2080	.154[b]	3.25[b]

[a] Predicted from groups 2 and 4.
[b] Predicted from groups 1 and 3.

This completes our analysis of the cue-additivity postulate relating learning rate to the relative weight of the relevant cues. The evidence is uniformly supportive to the postulate. The postulate has the additional advantage of permitting one to infer, via estimates of learning rates, something about the attentional or perceptual characteristics of the relevant attributes in the complex stimulus display.

3. The assumption of "single-cue" solutions

Implicit in the preceding application of the theory is the assumption that, upon solving the problem, the subject is attending to only one of the attributes of the stimulus pattern. Other available attributes (relevant or irrelevant) bear no essential relation to the subject's response tendencies. These nonessential attributes could be modified or deleted without seriously affecting the subject's performance to the selected relevant attribute.

We now report two results that favor this "one critical attribute" position. The first of these comes from an unpublished experiment by Bower and Wilkensen. Fifteen college students learned a two-category problem using three-letter consonant clusters as stimuli. The patterns were constructed according to the scheme outlined in Table 1. The letter in the leftmost position was P or M; F or Y was in the middle position; and G or W was in the rightmost position. The eight possible patterns were presented repeatedly on a memory drum to the subjects who classified the patterns as X's or

O's. The correct classification depended on the middle letter: if it were *F*, the answer was *O*; if it were *Y*, the answer was *X*.

After reaching a criterion of one perfect trial, the subjects were tested (without information feedback) on the 6 single letters (*P*, *M*, *F*, *Y*, *G*, and *W*), 12 pairs of letters (*PF*, *PY*, *MF*, *MY*, *PG*, *PW*, etc.) and the 8 three-letter clusters that were used during training. The percentage of correct responses to the three-letter clusters was .918 on the test trials, so there was some retention loss from the end of training to the test series. The question of interest is whether deletion of irrelevant letters results in any appreciable loss of performance on the relevant letters. The answer to this is negative. The probability of a correct response to the single letters *F* and *Y* was .890; the probability when a second, irrelevant letter was added (e.g., *PF*, *YW*, etc.) was .883. Neither probability reveals any appreciable loss below the .918 probability observed on tests to the three-letter patterns. Responses to only irrelevant letters, either singly or in pairs, were about evenly divided between *X*'s (.47) and *O*'s (.53), as was expected. We believe that this finding rules out the possibility that subjects learn the two-category problem by associating responses with complete patterns of stimulation including both relevant and irrelevant cues. The results are consistent with the assumption that the subject comes to attend to a single (relevant) cue. This cue is the sole discriminative stimulus for the subject, and his performance on it is essentially unaltered by deleting or altering the customary "background" of irrelevant stimulation.

The theory also leads to definite predictions about the effect of the partial validity of an irrelevant cue upon response probabilities when that cue is tested following concept learning. Remember that, prior to solving, the subject is viewed as selecting and conditioning values of the irrelevant attributes to the reinforced responses. After the subject solves, these conditional connections are left intact because irrelevant cues are no longer selected. Thus, a post-learning test with single irrelevant cues will reveal the conditioning (if any) of these cues that occurred before the subject solved on the relevant attribute. Suppose we let the phrase "reinforcement ratio" represent the frequency of A_1 to A_2 reinforcements occurring over a series of presentations of a value of an irrelevant dimension. Then the theory permits the inference that the probability that A_1 rather than A_2 is conditioned to a value of an irrelevant dimension comes to match the reinforcement ratio. For example, suppose color is relevant, form (circle or triangle) is irrelevant; suppose further that four out of five circular patterns presented are red and response A_1 is reinforced to red. For such conditions, the model predicts that a subsequent test with circle alone (without color) would yield an A_1 probability of .80. Results of this kind have been reported by Binder and Feldman (1960); they and Estes (1959b) have related these results to the postulates of stimulus-sampling theory. Because of the special role of attention in our formulation, we expect that the critical period for conditioning irrelevant

cues is the trials before solution (and consistent selection of the relevant cue). Thus, we would predict that a reinforcement differential to an irrelevant cue initiated after learning the relevant cue would not affect response probability on a later test to that irrelevant cue. This prediction has not yet been tested.

A second line of evidence giving partial support to the "one critical attribute" position comes from some transfer results reported by Trabasso (1961). The experiment and results will be reported only schematically here, but the experiment involved the redundant-cue groups 5 and 7 in Table 6. After learning with angle-plus-angle-color as redundant relevant cues (or angle-plus-flower-color), subjects were transferred to a problem in which the colors were removed but the angle size was retained as the relevant cue. The task for the theory is to predict performance on transfer. Suppose the two relevant cues are a and b. Their weights, w_a and w_b, can be assessed by calibration control groups (e.g., groups 2 and 4 in Table 6). Then the assumption of a single-cue solution implies that when a and b are redundant and relevant, a proportion $w_a/w_a + w_b$ of the subjects will solve during training using cue a (and not cue b), whereas a proportion $w_b/w_a + w_b$ of the subjects will solve using cue b. If all subjects are then transferred to the problem with cue b removed and only cue a relevant, the predictions are (1) that the proportion $w_a/w_a + w_b$ who solved initially on cue a will show perfect transfer, and (2) that the remaining proportion $w_b/w_a + w_b$ of the subjects who solved on cue b will show no transfer. Trabasso's data gave some support for these transfer predictions. The predicted average errors on transfer was too low for group 7, but was very close for condition 5. The predicted number of subjects showing errorless transfer was nearly exact in both cases. This evidence indicates that human subjects solve on the basis of a single relevant cue. The effect of adding more relevant cues is to increase the likelihood that any cue they choose for testing will lead to a solution.

The implication of these data is that when cues a and b are relevant, the subject actually learns about only one of these cues; when he solves on one cue, he ignores all other cues that may be redundant and relevant. We do not believe that this will be a general result, and the general properties of our model for acquisition depend in no way upon how the specific issues are decided via transfer tests. Suppose that the subject learns initially on a very difficult cue, a (e.g., its two values are barely discriminable). After he solves we make a second (previously irrelevant) cue b redundant and relevant, and suppose that cue b is extremely salient with highly discriminable values. Under such circumstances the subject might switch over to using the new cue b since the discrimination would then be easier for him. Hughes and North (1959) have reported some positive transfer when using rats in a discrimination experiment of this design. We suspect also that no transfer would be obtained to the added, second redundant cue if it were a more difficult cue than the first cue on which the subject solved the initial problem. Different

hypotheses of this kind could be easily tested using the deleted-cue procedure of the Bower-Wilkensen study.

4. The effective learning event in concept-identification experiments

The formal representation of the theory, outlined in the preceding pages, is a two-state Markov chain, with associated response probabilities p and 1. *On each reinforced trial* there is a constant probability, c, that a subject in the initial state will move into the terminal state. We now discuss the assumption that the subject may learn on any trial. There is some reason for thinking that the assumption is wrong, at least for the stimulus-selection process. Specifically, we believe that the critical informative events during stimulus selection occur only on those trials on which the subject makes an error. There are various factors that incline us to this view. If the subject is selecting and testing hypotheses, the only signal to indicate that he is on the wrong track occurs when he makes an error. When he is correct, he has no reason to discard the hypothesis he is using at the moment. In this sense, correct-response trials are simply dead-weights that retard the subject's selection of the correct hypothesis. The differential effect of correct and incorrect response trials on the subject's behavior is obvious to the observer. During a string of correct responses the subject relaxes, gives his responses quickly, and does not pay much attention to the stimulus display after he receives the information indicating his response was correct. In contrast, following an error, the subject is alert and searches through the attributes on the stimulus display, and his response latencies on the next trial or two are slower than his response latencies following correct-response trials.

To apply this modified model to our data, prediction equations will be derived. Let ε represent the probability of learning on an error trial; the probability of learning on a correct response trial is assumed to be zero. We begin by writing out the matrix of transition probabilities for the Markov chain. Let C represent the terminal, conditioned state. For convenience in analysis, the unlearned state is distinguished according to the subject's response: state E, which the subject is in just before he makes an error, and state S, which the subject is in just before he makes a correct response (success). The probability of a correct response when the subject is in state C, S, or E is 1, 1, or 0, respectively. If the subject is in state S, he cannot learn; with probability p he will be in state S on the next trial and with probability q he goes into state E for the next trial. From state E, the subject learns with probability ε; but with probability $1 - \varepsilon$, he does not learn, and on the next trial he will be in state E or state S with probabilities q and p, respectively. The state-to-state transition probabilities are summarized in the matrix on page 52. To the right of the matrix are indicated the probabilities that the subject begins on trial 1 in each of the three states.

We begin derivations by obtaining the distribution of total errors before learning. Call this random variable T. From (12), it is apparent that state C

(12)

State on trial n	State on trial $n+1$: C	E	S	Prob. of state on trial 1
C	1	0	0	0
E	ε	$q(1-\varepsilon)$	$p(1-\varepsilon)$	q
S	0	q	p	p

can be entered only through state E. Hence, at least one error must occur. The probability that any particular error is the last one is ε. From these considerations the probability of exactly k errors is

$$P(T = k) = \varepsilon(1 - \varepsilon)^{k-1} \qquad (k = 1, 2, 3, \cdots). \tag{13}$$

The mean and variance of total errors are

$$E(T) = \frac{1}{\varepsilon} = u_1,$$
$$\text{var } (T) = \frac{1 - \varepsilon}{\varepsilon^2} = u_1^2 - u_1. \tag{14}$$

Two parenthetical comments concerning the total error distribution are in order. The first comment is that this model gives the same cue-additivity results obtained previously because the expected errors remain a simple reciprocal function of the learning parameter. Thus, assuming that $\varepsilon = \theta r$, all previous numerical predictions about cue-additivity experiments would follow from Eq. (14) for mean errors. The second comment is that Eq. (13) states that no subjects will solve with zero errors; if a subject can learn only on errors, then everyone who learns must make at least one error. This implication is contrary to fact: there was one subject in our experiments who met criterion without making any errors. To correct for this, one might assume that a proportion ε of the subjects start with the correct hypothesis. This changes the state probabilities on trial 1 to read ε, $q(1 - \varepsilon)$, and $p(1 - \varepsilon)$ instead of 0, q, and p for states C, E, and S. Since we have observed only a single case of zero errors, we do not make this correction in the analysis which follows.

For derivations of most results, we need to know the probability that a subject will be in state S or E at the beginning of trial n, given that he starts off on trial 1 in states S or E with probabilities p and q. Let $w_{S,n}$ and $w_{E,n}$ represent these probabilities. Recursive expressions for them are:

$$w_{E,n+1} = w_{E,n}\, q(1 - \varepsilon) + w_{S,n} \cdot q,$$
$$w_{S,n+1} = w_{E,n}\, p(1 - \varepsilon) + w_{S,n} \cdot p. \tag{15}$$

These equations are the same except that p and q are interchanged. If the

top equation is multiplied by p and the bottom one by q, the right-hand sides of the two equations are identical. Hence for all trials, $pw_{E,n} = qw_{S,n}$. If we use this identity to substitute for $w_{S,n}$ in the top equation, the equation reads

$$w_{E,n+1} = w_{E,n}q(1 - \varepsilon) + w_{E,n}p \tag{16}$$
$$= (1 - q\varepsilon)w_{E,n}.$$

Equation (16) is a linear difference equation that can be solved to yield

$$w_{E,n} = (1 - q\varepsilon)^{n-1}w_{E,1} \tag{17}$$
$$= (1 - q\varepsilon)^{n-1}q,$$

where in the second line the initial condition, $w_{E,1} = q$, has been substituted. Because of the identity $pw_{E,n} = qw_{S,n}$, the expression for $w_{S,n}$ is

$$w_{S,n} = (1 - q\varepsilon)^{n-1}p. \tag{18}$$

Errors occur only when the subject is in state E. Hence, Eq. (17) for $w_{E,n}$ gives the average probability of an error on trial n. Summation of $w_{E,n}$ over all trials leads to the expression $1/\varepsilon$, the average total errors.

We shall now derive the probability distribution of the trial of the last error. The result is direct: the probability that the last error occurs on trial n is just $w_{E,n}\varepsilon$—that is, $w_{E,n}$ is the probability that an error occurs on trial n, and ε is the likelihood that the subject learns on that error trial. If we let n' represent the trial of the last error, the distribution is

$$\Pr\{n' = k\} = q\varepsilon(1 - q\varepsilon)^{k-1}, \tag{19}$$

having mean and variance equal to

$$E(n') = \frac{1}{q\varepsilon}, \tag{20}$$
$$\operatorname{var}(n') = \frac{1 - q\varepsilon}{(q\varepsilon)^2}.$$

The mean trial of the last error is just $1/q$ times the total errors. If q is one-half and $T = 10$, then $E(n') = 20$, which seems reasonable.

An experimental result reported by Gormezano and Grant (1958) supports this relation between average errors, q, and the trial of the last error. In a card-sorting concept-identification task, these investigators experimentally varied the probability that a subject would be correct if he were sorting on the basis of certain irrelevant hypotheses. This probability can be varied by manipulating the frequency of presentation of the 2^n patterns so that some irrelevant cues correlate better than chance with the correct answer. Their finding was that variations in the amount of partial validity of irrelevant cues (our p) had no influence on the average errors before subjects learned. The main effect of such variations was to increase the number of correct responses before the last error and, hence, to increase the average trial of the last error. We have obtained the same results in a pilot study using single-stimulus presentation and binary (verbal) classification. These

results are qualitatively consistent with Eq. (20) relating $E(T)$ and $E(n')$. An interesting problem for future work is to develop a microtheory that makes exact predictions of p [and hence $E(n')$] as the partial validities of varying numbers of irrelevant cues are manipulated.

Returning to the analysis of the model, we now derive the distribution of the number of errors between the kth and $(k+1)$th success. Let J_k represent this random variable. The distribution of J_k will depend on the probability that the kth success occurred in state S rather than in the conditioned state C. Let g_k represent the probability that the kth success occurred by chance, in state S. Using g_k, we may write the distribution of J_k as

$$\Pr\{J_k = i\} = \begin{cases} 1 - qg_k & \text{for} \quad i = 0, \\ g_k q(1-\alpha)\alpha^{i-1} & \text{for} \quad i \geq 1, \end{cases} \tag{21}$$

where $\alpha = q(1-\varepsilon)$. To have i errors between kth success and the next one, it must be that the kth success occurred in state S, with probability g_k; that the subject moves to state E with probability q and stays there for $i-1$ trials, with probability $q(1-\varepsilon)$ on each trial; and that after the ith error, a success occurs on the next trial, with probability $\varepsilon + (1-\varepsilon)p$. The quantity $\alpha = q(1-\varepsilon)$ represents a basic observable quantity in the data; it is the conditional probability of an error, given an error on the prior trial. The mean and variance of J_k are

$$\begin{aligned} E(J_k) &= \frac{q}{1-\alpha} g_k, \\ \operatorname{var}(J_k) &= \frac{qg_k}{(1-\alpha)^2}[1 + \alpha - qg_k]. \end{aligned} \tag{22}$$

The distribution of S_k, the number of errors between the kth and $(k+2)$th success, has been obtained and is presented here for completeness. The result is

$$\Pr\{S_k = j\} = \begin{cases} (1 - g_k(1-p^2) & \text{for} \quad j = 0, \\ g_k[q\varepsilon + p^2(1-\varepsilon) + jp(1-\alpha)]\alpha^{j-1} & \text{for} \quad j \geq 1. \end{cases} \tag{23}$$

The mean and second raw moments of S_k are

$$\begin{aligned} E(S_k) &= \frac{qg_k}{(1-\alpha)}\left[1 + \frac{p}{1-\alpha}\right] = E(J_k) + E(J_{k+1}), \\ E(S_k^2) &= \frac{g_k pq}{(1-\alpha)^3}\{2\alpha + (1+\alpha)(2-\varepsilon)\}. \end{aligned} \tag{24}$$

To complete matters, an expression for g_k, the probability that the kth success occurs with the subject in state S, is required. A recursion on g_k is

$$\begin{aligned} g_{k+1} &= g_k[p + q(1-\varepsilon)p + q^2(1-\varepsilon)^2 p + \cdots] \\ &= g_k p \sum_{i=0}^{\infty} \alpha^i = \frac{p}{1-\alpha} g_k. \end{aligned} \tag{25}$$

The difference equation in (25) may be solved to obtain

$$g_k = g_1\left(\frac{p}{1-\alpha}\right)^{k-1}. \tag{26a}$$

Finally, the value of g_1, the likelihood that the first success occurs in state S, is $p/1-\alpha$, obtained from Eq. (25) by setting $g_k = 1$. The final equation for g_k is

$$g_k = \left(\frac{p}{1-\alpha}\right)^k = \left(\frac{p}{p+q\varepsilon}\right)^k \qquad (k = 1, 2, \cdots). \tag{26b}$$

The value of g_k decreases exponentially at a rate depending primarily on ε; the larger ε is, the less likely it is that the kth success occurred by chance. With the general expression for g_k, the average value of J_k may be rewritten as

$$E(J_k) = \frac{q}{1-\alpha}\left(\frac{p}{1-\alpha}\right)^k. \tag{27}$$

J_k has been defined as the number of errors between the kth and $(k+1)$th success; J_0 is the number of errors before the first success. If we set $g_0 = 1$, then Eq. (21) gives the distribution of errors before the first success and Eq. (23) gives the distribution of errors before the second success.

If errors between adjacent successes are cumulated, the limit of this cumulative function should be the total number of errors. Define F_k as the cumulative errors before the kth success. Then

$$E(F_k) = \sum_{i=0}^{k-1} E(J_k). \tag{28}$$

Substitution into Eq. (28) of the expression for $E(J_k)$ from Eq. (27), recalling that $g_0 = 1$, leads to the following result:

$$E(F_k) = \frac{1}{\varepsilon}\left[1 - \left(\frac{p}{1-\alpha}\right)^k\right]. \tag{29}$$

The limit of this expression as k becomes large is $1/\varepsilon$, the expected total number of errors.

The quantity $1 - g_k$ is the probability that no errors follow the kth success, since $1 - g_k$ is the likelihood that the kth success occurs in state C. One can use this observation to derive the probability distribution of the number of successes before the last error. Let Z represent this random variable. The quantity $1 - g_k$ is the cumulant distribution of Z, i.e., $1 - g_k = P(Z \leq k-1)$. The probability that $Z = k$ may be obtained by taking differences between adjacent values of the cumulant distribution; i.e.,

$$\begin{aligned} \Pr\{Z = k\} &= (1 - g_{k+1}) - (1 - g_k) = \frac{q\varepsilon}{1-\alpha}\left[\frac{p}{1-\alpha}\right]^k \\ &= (1 - g_1)g_1^k \qquad (k = 0, 1, 2, \cdots). \end{aligned} \tag{30}$$

The mean and variance of the number of successes before the last error are

$$E(Z) = \frac{g_1}{1 - g_1} = \frac{p}{q\varepsilon}, \tag{31}$$

$$\text{var}\,(Z) = \frac{g_1}{(1 - g_1)^2} = \frac{(1 - \alpha)p}{(q\varepsilon)^2}.$$

The sum of errors and successes before the last error equals the trial of the last error. In equation form,

$$E(n') = E(T) + E(Z), \qquad \frac{1}{q\varepsilon} = \frac{1}{\varepsilon} + \frac{p}{q\varepsilon}. \tag{32}$$

Aegebraic manipulation of the right side of Eq. (32) establishes the identity. Bacause of this identity, $E(Z)$ is not an independent prediction once $E(T)$ and $E(n')$ are known or predicted. However, independent free predictions can be made of the distribution and variance of Z specified by Eqs. (30) and (31).

Another random variable that carries information about the response sequences is the distribution of success runs between adjacent errors. The distribution of this random variable depends only on p. The learning parameter ε does not enter because the random variable is defined with the requirement that another error will eventually follow the error on the reference trial. Let H represent the number of successes between adjacent errors. Its distribution is

$$\Pr\{H = k\} = qp^k \qquad (k = 0, 1, 2, \cdots), \tag{33}$$

having mean and variance

$$E(H) = \frac{p}{q}, \quad \text{var}\,(H) = \frac{p}{q^2}. \tag{34}$$

Equation (33) also gives the distribution of the number of successes before the first error. In experimental tests of the distribution specified in Eq. (33), observations before the first error and between subsequent adjacent errors are pooled.

The remaining statistics to be considered are error runs, alternations, and autocorrelations, which summarize various sequential features of the response series. First, we obtain an expression for the average number of alternations of success and failures that occur before a subject learns. The probability of obtaining an alternation between trials n and $n + 1$ is

$$A_n = w_{E,n}(1 - \alpha) + w_{S,n}q.$$

Tne expected number of alternations over all trials n is

$$A = \sum_{n=1}^{\infty} A_n = \frac{p + 1 - \alpha}{\varepsilon}. \tag{35}$$

Consider next the average number of times pairs of errors occur k trials apart. Call this statistic c_k. For example, c_2 would be the mean number of error-pairs occurring across one intervening trial; that is, pairs of errors on trials n and $n + 2$ are summed over all trials. The probability of counting a pair of errors k trials apart when the leading error occurs on trial n is

$$c_{k,n} = w_{E,n}\, q(1 - \varepsilon)(1 - q\varepsilon)^{k-1}.$$

The mean total number of error-pairs k trials apart is obtained by summing $w_{E,n}$ in this equation over all trials:

$$\begin{aligned} c_k &= \sum_{n=1}^{\infty} (1 - q\varepsilon)^{n-1} q\, q(1 - \varepsilon)(1 - q\varepsilon)^{k-1} \\ &= q\frac{(1 - \varepsilon)}{\varepsilon}(1 - q\varepsilon)^{k-1} \\ &= q(u_1 - 1)(1 - q\varepsilon)^{k-1}. \end{aligned} \tag{36}$$

A similar statistic may be obtained for the average number of times a success is followed k trials later by an error. Call this statistic d_k. Its over-all trial sum is

$$d_k = pu_1(1 - q\varepsilon)^{k-1}. \tag{37}$$

Equations (36) and (37) establish that the values of the statistics c_k and d_k should decrease by a fraction $1 - q\varepsilon$ as k increases. If ε is small (.10 or less), as it is in our experiments, then $1 - q\varepsilon$ is .95 or higher and the expected decrease with k in c_k and d_k is relatively small, especially so compared with the absolute values of c_1 and d_1. Thus, for small ε, these particular sequential statistics are not very informative since their variation with k is small with respect to the sampling variability in the c_k, d_k statistics. This observation should be kept in mind when predictions of c_k are considered in later sections.

The final statistics to be considered are those referring to runs of errors. Let r_j represent the mean number of runs of exactly j errors and let $R = \Sigma_j r_j$ represent the mean number of error runs of any length. These r_j values are obtained by first deriving the value of u_j, the expected number of j-tuples of errors (i.e., an uninterrupted string of j errors). Once these u_j are obtained, R and r_j may be obtained by the relations (cf. Bush, 1959):

$$\begin{aligned} R &= u_1 - u_2, \\ r_j &= u_j - 2u_{j+1} + u_{j+2}. \end{aligned} \tag{38}$$

The probability of an uninterrupted string of j errors from trial n through trial $n + j - 1$ inclusive is

$$u_{j,n} = w_{E,n}[q(1 - \varepsilon)]^{j-1} = w_{E,n}\alpha^{j-1}.$$

The expected number of j-tuples of errors summed over all trials is

$$u_j = \frac{\alpha^{j-1}}{\varepsilon} = u_1\alpha^{j-1}. \tag{39}$$

From Eqs. (38) and (39), we obtain

$$R = u_1(1-\alpha), \qquad r_j = u_1(1-\alpha)^2\alpha^{j-1}. \tag{40}$$

This completes the mathematical analysis of the model. It is apparent that a number of predictions for a given set of data can be derived easily from the axioms of the model.

5. Estimation of parameters

The stochastic process identified with the model is a function of two parameters, p and ε. In specific applications these parameters are estimated from the data. In this section are derived estimators of p and ε that have the property that they maximize the likelihood of the observed data. The material in this section stems from an excellent paper by Restle (1961) on estimation techniques for all-or-none learning models.

We begin by writing the likelihood of the sequence of responses obtained from subject i. Let x_{in} be his response random variable, taking the value 1 or 0 accordingly as he makes an error or correct response on trial n. A sequence of x_i's is obtained from this subject and the probability of this sequence is

$$l_i = f(x_{i1}, x_{i2}, \cdots) = p^{z_i}q^{t_i}(1-\varepsilon)^{t_i-1}\varepsilon. \tag{41}$$

The right side of Eq. (41) summarizes the subject's response series by two numbers: t_i, his total number of errors, and z_i, the number of successes before his last error. Because responses prior to the last error are independent, all possible sequences of z_i successes and t_i errors have the same probability, and we may ignore the actual order in which the successes and failures occurred. Since order is immaterial, the statistics z_i and t_i are sufficient statistics; that is, these numbers summarize all the information in the response series relevant to the parameters of the stochastic process.

Now suppose that there are N subjects, all of whom meet the learning criterion. The joint likelihood function for the N subjects is the product of the individual likelihood functions given by Eq. (41). In taking products, the exponents z_i and t_i are summed over subjects. Let $Z = \Sigma_{i=1}^{n} z_i$ and let $T = \Sigma_{i=1}^{n} t_i$. The joint likelihood function is

$$L = p^Z(1-p)^T(1-\varepsilon)^{T-N}\varepsilon^N. \tag{42}$$

Because z_i and t_i are sufficient statistics for each individual sequence, Z and T are also sufficient statistics for the pooled data from N subjects. We want estimates of p and ε that maximize L. Because Eq. (42) can be factored into the product of a function depending only on p and a function depending only on ε, L can be maximized with respect to one parameter without regard to the value of the other parameter.

It is convenient to find the maxima of $\log L$ rather than of L itself. Because $\log L$ is a strictly monotonic-increasing function of L, the maxima of $\log L$

will also be the maximal points of L. Thus, the beginning expression is

$$\log L = Z\log p + T\log(1 - p) + (T - N)\log(1 - \varepsilon) + N\log\varepsilon. \tag{43}$$

The estimate for ε is obtained by setting equal to zero the partial derivative of $\log L$ with respect to ε, namely,

$$\frac{\partial \log L}{\partial \varepsilon} = \frac{N}{\varepsilon} - \frac{(T - N)}{1 - \varepsilon} = 0. \tag{44}$$

Solution of Eq. (44) for ε yields the desired maximum likelihood estimate:

$$\hat{\varepsilon} = \frac{N}{T} = \frac{1}{\bar{T}}. \tag{45}$$

This estimate of ε is an intuitively natural one. We saw in Eq. (14) that the expected total errors is $1/\varepsilon$. The estimate in Eq. (45) simply solves this equation for ε in terms of the observed $\bar{T}$.

The variance of the maximum likelihood estimate is given by

$$\text{var}(\hat{\varepsilon}) = \frac{-1}{E\left[\frac{\partial^2 \log L}{\partial \varepsilon^2}\right]}. \tag{46}$$

The negative of the second partial derivative of $\log L$ with respect to ε is

$$-\frac{\partial^2 \log L}{\partial \varepsilon^2} = \frac{N}{\varepsilon^2} + \frac{(T - N)}{(1 - \varepsilon)^2}. \tag{47}$$

The only random variable in Eq. (47) is T, the total number of errors made by N subjects. The expected value of T is N/ε. Substituting this into Eq. (47), simplifying, and taking the reciprocal in accord with Eq. (46), we obtain the result

$$\text{var}(\hat{\varepsilon}) = \frac{\varepsilon^2(1 - \varepsilon)}{N}. \tag{48}$$

Substituting into Eq. (48) the estimate $\varepsilon = N/T$, we obtain

$$\text{var}(\hat{\varepsilon}) = \frac{N(T - N)}{T^3}. \tag{49}$$

To illustrate, for the experiment reported in Table 4, $T = 304$ and $N = 25$; hence, $\hat{\varepsilon} = .0823$ and $\sigma(\hat{\varepsilon}) = .0158$. The maximum likelihood estimate of ε provides meaningful and powerful statistical tests of hypotheses stated in terms of the ε values. In a discussion of subsequent experiments, these properties of the ε estimates are used for making statistical decisions.

We now derive a maximum likelihood estimate for p. Referring to Eq. (43) for $\log L$, we obtain the first derivative of $\log L$ with respect to p:

$$\frac{\partial \log L}{\partial p} = \frac{Z}{p} - \frac{T}{1 - p} = 0. \tag{50}$$

The solution for p from Eq. (50) is

$$\hat{p} = \frac{Z}{Z + T} = \frac{\bar{Z}}{\bar{Z} + \bar{T}} = \frac{\bar{Z}}{\bar{n}'}. \tag{51}$$

The variance of this estimate of p is obtained by methods analogous to those in Eqs. (46) through (49). The result is

$$\text{var}(\hat{p}) = \frac{pq^2}{T} = \frac{ZT}{(Z+T)^3}. \tag{52}$$

To illustrate, in Table 4 the relevant values are $T = 304$ and $Z = 338$. Hence, $\hat{p} = .526$ and $\sigma(\hat{p}) = .0196$.

In the experimental section, statistical hypotheses of the form $\varepsilon_1 = \varepsilon_2$ are tested where ε_1 and ε_2 are estimated from groups of subjects given different treatments. The alternative hypothesis is $\varepsilon_1 \neq \varepsilon_2$. An elegant test of the null hypothesis is obtained by taking the ratio of likelihoods as follows:

$$\lambda = \frac{\max_{\varepsilon_c} L(T_1, T_2; \varepsilon_c)}{\max_{\varepsilon_1,\varepsilon_2} L(T_1; \varepsilon_1) L(T_2; \varepsilon_2)} = \frac{(1-\varepsilon_c)^{T_1+T_2-N_1-N_2}\varepsilon^{N_1+N_2}}{(1-\varepsilon_1)^{T_1-N_1}\varepsilon_1^{N_1}(1-\varepsilon_2)^{T_2-N_2}\varepsilon_2^{N_2}}. \tag{53}$$

The numerator of this ratio is the maximum likelihood of the data from the two groups under the restriction that ε_1 and ε_2 are equal to the common estimate, ε_c. The denominator is the likelihood of the data when ε_1 and ε_2 are chosen separately, without restriction, to maximize the likelihoods for the two sets of data. The ratio λ is a fraction between zero and one. If λ is close to unity, then the likelihood of the data based on the common ε_c estimate is about as large as the likelihood when ε_1 and ε_2 are chosen freely; accordingly, when λ is close to 1, we accept the null hypothesis that $\varepsilon_1 = \varepsilon_2$. Conversely, if ε_1 and ε_2 differ markedly, then λ will be close to zero and we reject the null hypothesis. A cutting point, c, is chosen in the unit interval so that if $\lambda \geq c$ the null hypothesis is accepted, and if $\lambda < c$ the null hypothesis is rejected. The choice of the cutting point is aided by the fact that $y = -2 \ln \lambda$ has a χ^2 distribution. The degrees of freedom for y are equal to the difference in the number of parameters estimated in the denominator (two here) and the number estimated in the numerator (one here). Accordingly, our hypotheses are tested by calculating $-2 \ln \lambda$ and then using tables of the χ^2 distribution.

6. Comparisons of model with data from new experiments

Having completed the mathematical analysis of the model, we now consider its various applications to experimental data. The experiments will be discussed in the chronological order in which they were done. First, we wanted to reach some kind of convincing quantitative decision between the two rival assumptions, viz., that learning occurs only on error trials, or that learning can occur on either correct-response or error trials. Our first strategy in trying to decide this issue was to compare the goodness of fit of the two models to details of data from ordinary two response experiments such as the one displayed previously in Table 4. After comparing the models on several sets of data, we concluded that this was a poor decision strategy. The problem was that fine-grain details of such data were fitted about equally well by both

models. Quantitative predictions of the two models differed only slightly and the predictions of one model were not consistently closer to the data statistics. This lack of differentiation means that most of the fine-grain statistics we have calculated are primarily sensitive to the existence of a stationary Bernoulli trials process prior to the last error. Both models assume such a Bernoulli process prior to the last error. Since this property characterizes a large portion of the sequence of responses obtained from a subject, we were actually trying to differentiate the models according to their assumptions about the remaining portion of the data. This is surely a losing strategy for testing theoretical assumptions in the psychology of learning.

The alternative strategy we adopted was to arrange new experimental conditions so that the different assumptions led to sizable differences in the expected outcome of the experiment. We carried out two experiments that differentiate clearly between these different learning assumptions (i.e., learning on all trials vs. learning only on error trials). The first of these experiments is discussed now; the second will be discussed later in another context.

Our first experiment varied the chance probability of an error by manipulating the number of response alternatives. This manipulation achieves the same effect on p as did Gormezano and Grant's (1958) manipulation of the partial validity of irrelevant cues. The stimulus patterns consisted of outlined geometrical figures with three variable dimensions: color, shape, and the angle of a line slicing through the figure. There were four values of each dimension. The colors were red, green, blue, and brown; the shapes were circles, squares, triangles, and hexagons; and the angles were 0, 45, 90, and 135 degrees. Color was the relevant attribute. For one group, consisting of 22 college students, there were four responses (numbers 1, 2, 3, and 4), with each response paired off uniquely with one of the four values of the relevant color dimension. For the second group of 22 subjects, only two responses (1 or 2) were used, with each response paired with two of the color values. For example, the assignments might be that red and blue were 1's, and that green and brown were 2's. Each subject continued training until he met a criterion of 16 consecutive correct responses.

In all of our experiments, the instructions to the subjects were essentially the same. The subjects were told that their task was to learn to classify a set of stimulus patterns on the basis of a simple principle. The stimulus patterns were to be classified into two (in one instance, four) classes, and all patterns belonging to a class had some common defining property. This instructional procedure differs from that used by other investigators of concept learning (e.g., Bourne, 1957; Bruner *et al.*, 1956). In our studies, the subject was not given detailed information regarding the available stimulus attributes and their possible classifications. Therefore, the characteristics of the stimuli were evaluated by the subject as they appeared to him. To orient the subject to the task, we frequently resorted to pretraining him on an easier but unrelated two-choice problem.

The a priori probability of guessing correctly for the two-response

subjects is $\frac{1}{2}$; for the four-response subjects, the probability of a correct guess is much less than $\frac{1}{2}$, the lower limit for unintelligent guessing being $\frac{1}{4}$. We are not concerned here with the fact that the p-value for the four-response group be $\frac{1}{4}$; as a matter of fact, it was nearer $\frac{1}{3}$. All that we expected was that the p for the four-response subjects would be less than that for the two-response subjects.

Given that there are differences in the p values for the two experimental conditions, the two assumptions make different predictions. The assumption that learning occurs only on error trials predicts that the average total number of errors to criterion will be the same $(1/\varepsilon)$ for both conditions but that the conditions will differ on the average trial of the last error $(1/q\varepsilon)$. If p is equal to $\frac{1}{2}$ and to $\frac{1}{3}$ for the two experimental conditions, then the average trial of the last error will be about one and a half times larger for the two-response subjects. The alternative assumption that learning occurs on all trials predicts that average errors, q/c, will be about 1.5 times larger for the four-response subjects, and that the average trial of the last error,

$$\frac{1/c}{1 + c(p/q)},$$

will be only slightly affected by p. Bower (1962a) has reported paired-associate results supporting predictions of this latter type. However, in paired-associate experiments it may be reasonable to suppose that the subject learns on all trials because he can recognize when a correct response is only a guess. The paired-associate data is not relevant to deciding the validity of the assumption regarding the stimulus-selection process in concept-identification learning.

The outcome of the experiment favored the assumption that stimulus-selection learning takes place only on error trials. The average total errors for the two conditions were similar, 13.36 for the two-choice subjects and 12.41 for the four-choice subjects. The ε estimates were $\hat{\varepsilon}_1 = .080$ for the four-choice group, and $\hat{\varepsilon}_2 = .075$ for the two-choice group. When both groups are pooled, the common estimate of ε_c is .078. The null hypothesis that $\varepsilon_1 = \varepsilon_2 = \varepsilon_c$ is tested by the likelihood ratio in Eq. (53). Substituting the estimates for ε_1, ε_2, and ε_c into Eq. (53), we obtain the value $y = .19$. With one degree of freedom, a chi-square value equal to or greater than .19 would be expected to occur more than 60 per cent of the time if the hypothesis that $\varepsilon_1 = \varepsilon_2$ were true. Hence, on the basis of this test we accept the hypothesis that $\varepsilon_1 = \varepsilon_2$. Thus, variation in the number of response alternatives resulted in no significant differences in the learning rate or in the total errors.

In accord with the error-learning assumption, the experimental groups differed in their average trial of the last error. The mean for the two-response group was approximately 1.5 times the mean for the four-response group (26.36/18.32 = 1.44). These differences, of course, are interpreted as due to differences in the p parameter. Using Eq. (51), the estimate we obtained for

the two-choice group was $\hat{p}_1 = .493$, and the estimate for the four-choice group was $\hat{p}_2 = .323$. When the two groups were pooled, the common estimate of $\hat{p}_c$ was .423. The hypothesis that $p_1 = p_2 = p_c$ was tested by the likelihood ratio:

$$\lambda = \frac{p_c^{Z_1+Z_2}(1-p_c)^{T_1+T_2}}{p_1^{Z_1}(1-p_1)^{T_1}p_2^{Z_2}(1-p_2)^{T_2}}.$$

After substituting the above estimates for the p_i and taking $-2 \ln \lambda$ of the quantity, we obtained 28.67 as the value of y. With one degree of freedom, a χ^2 at 28.64 or greater would occur less than 1 out of 1000 times if the null hypothesis were true. Hence, we reject the hypothesis that $p_1 = p_2$. Thus, the manipulation of the number of response alternatives produced large, significant differences in the value of p.

The response sequences prior to the last error were analyzed for the expected properties of an independent Bernoulli trials process. The analysis for stationarity of response probabilities prior to the last error was carried out in two ways. First, it was done in blocks of 10 trials, forward from trial 1, dropping subjects who meet criterion. For the two-response conditions, $\chi^2 = 2.49$ with 8 degrees of freedom, giving a $P > .95$; for the four-response condition, $\chi^2 = 10.20$ with 4 degrees of freedom, giving a $.05 > P > .02$. A second analysis for stationarity divided each subject's responses prior to his last error into quarters, then pooled across subjects. This Vincentized analysis is sensitive to small trends that might appear near the end of each individual's sequence. The χ^2 test, based on three degrees of freedom, gave a value of 1.16 ($P > .70$) for the two-response subjects and a value of 8.09 ($.05 > P > .02$) for the four-response subjects. Thus, the two-response data exhibit stationarity; the four-response data yield borderline significance. These four-response data are discussed in a more general context in the Appendix.

A second analysis was performed to test the hypothesis that successive

TABLE 7

TRANSITION FREQUENCIES FOR INDEPENDENCE TEST (TWO- AND FOUR-RESPONSE EXPERIMENT)

Two-Response Data:		Trial $n+1$ Success	Trial $n+1$ Failure	$P(S \mid X)$
Trial n	Success	132	140	.485
	Failure	137	129	.514
Four-Response Data:		**Trial $n+1$ Success**	**Trial $n+1$ Failure**	**$P(S \mid X)$**
Trial n	Success	45	71	.388
	Failure	82	161	.333

responses were statistically independent. The observed transition frequencies are given in Table 7. The χ^2 values obtained from these tables are 48. for the two-response data and .78 for the four-response data. With one degree of freedom, the values of .48 and .78 both give $P > .30$. Thus, in neither case is there sufficient cause to reject the null hypothesis that successive responses are statistically independent.

A statistic that depends only upon the Bernoulli-trials assumption (and not upon ε) is the distribution of the number of successes between adjacent errors. The probability of k successes between adjacent errors should be qp^k. The empirical and theoretical distributions of this random variable are shown in Fig. 4. The distribution on the left is for the two-response data, and the distribution on the right for the four-response data. Both empirical distributions correspond closely to the predicted values.

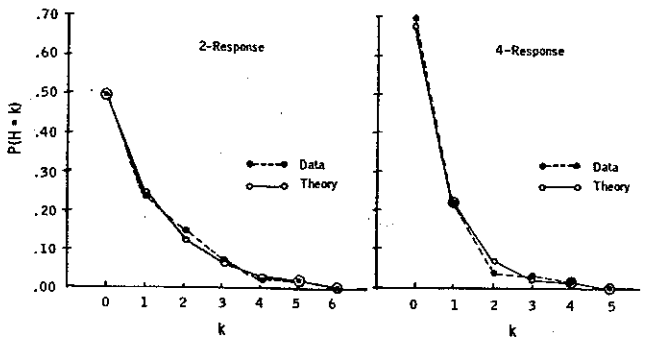


FIG. 4. Distribution of H, the number of successes between adjacent errors, for the two- and four-response groups.

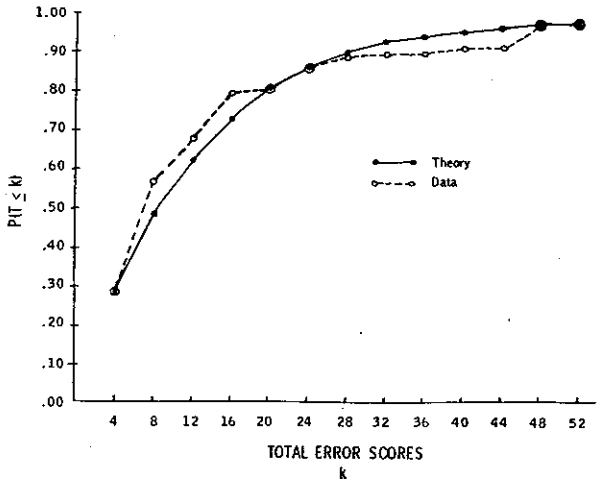


FIG. 5. Cumulative distribution of T, total errors, plotted in class intervals of four errors (two- and four-response data combined).

We now compare with data those predictions that are a function of the learning parameter. A critical prediction concerns the distribution of the total number of errors. Since the error distributions for the two groups are highly similar (they were expected to be identical), we pool the two groups of scores for this analysis. In Fig. 5 is plotted the predicted and observed cumulative distribution of errors. The theoretical function is obtained by setting $\varepsilon = .078$ in

$$\Pr\{T \leq k\} = 1 - (1 - \varepsilon)^k. \tag{54}$$

The goodness of fit of theoretical to empirical values was evaluated by the Kolmogorov-Smirnov one-sample test. The largest observed discrepancy was .09. According to tabled probabilities of such deviations (Siegel, 1956), a deviation of .09 or larger would be expected more than 20 per cent of the

TABLE 8

TWO- AND FOUR-RESPONSE EXPERIMENT: DETAILED PREDICTIONS

Statistic	Two-Response		Four-Response	
	Observed	Predicted	Observed	Predicted
Average errors	13.36	13.36	12.41	12.41
Standard deviation	16.83	12.82	9.74	11.89
Average trial of last error	26.36	26.67	18.32	18.25
Standard deviation	31.90	26.16	14.38	17.74
Average errors before first success	1.04	.93	3.18	1.79
Standard deviation	1.29	1.33	3.44	2.12
Average errors between first success and second success	.64	.87	1.68	1.54
Standard deviation	.66	1.28	1.96	1.80
Average successes before last error	13.00	13.36	5.91	5.89
Standard deviation	15.36	13.83	5.57	6.37
Runs of errors	7.23	7.19	4.73	4.69
Runs of length 1	3.73	3.87	1.95	1.77
2	1.95	1.79	1.18	1.10
3	.82	.83	.23	.68
4	.59	.38	.54	.42
Average error pairs				
1 trial apart	6.14	5.94	7.68	7.30
2 trials apart	5.45	5.71	7.54	6.90
3 trials apart	5.59	5.50	7.32	6.53
Alternations of success and failure	13.82	12.83	8.54	8.65
Probability of an error following an error	.458	.462	.619	.622

time with a sample of 44 cases. Hence, the test does not reject the fit of theoretical to observed values.

A list of point predictions is given in Table 8 for the two- and four-response conditions. In general, the model accurately predicts variances and sequential statistics. One poor prediction for the four-response data is that of the errors before the first success. This discrepancy will be discussed in the Appendix.

7. The reversal experiments

A series of three reversal experiments will be presented that support the notion of all-or-none learning for Type I concept problems. The first two experiments are similar; the second experiment was a replication of the first with different stimulus materials. The third experiment differs from the first two and will be discussed separately.

All three experiments test the same point in all-or-none theory, namely, that an error is a recurrent event in the subject's response protocol—that is, the error indicates that the subject is in the same state as he was at the beginning of the problem and that the intervening trials have had no particular effect upon his probability of solving. An equivalent form of this assumption is that the probability of learning after any given error is a constant, ε. For those subjects who make an error on trial n, the conditional distribution of the number of subsequent errors should be the same as the unconditional error distribution for all subjects from trial 1. Data confirming this prediction for paired-associate learning and for verbal-discrimination learning have been reported by Bower (1962a). Here the theorem is tested with two-response concept-identification learning.

In the first experiment, subjects were reinforced initially for the wrong classification. Before they had learned this, the response assignments were either reversed or were shifted to a different cue. The question is whether the initial wrong-way training retards learning on the final problem.

Readers familiar with the history of the continuity–noncontinuity issue will recognize the experiment as a familiar design in this area. Reversal experiments of this type have been performed by a number of investigators, notably those done by McCulloch and Pratt (1934), Krechevsky (1938), and Ehrenfreund (1948). The early experiments were reviewed by Blum and Blum (1949). Without exception, the experiments involved rats as subjects, simultaneous discriminations, and simple cues. On balance, the evidence tended to favor a continuity position supplemented by constructs such as peripheral receptor-orienting acts or observing responses. However, this body of evidence with nonarticulate, infra-human organisms may not be of particular relevance to human concept-identification learning carried out with successive discrimination training. Writing as a spokesman for the continuity position, Spence (1940) pointed out that the results obtained in the animal studies may not be directly relevant to human learning mediated by

complex symbolic mechanisms. Consequently, the existing literature on presolution reversal in animals did not discourage us from carrying out our experiments with human subjects. As will be seen, the resulting dividends justified the strategy.

The procedure for the three conditions in our experiment is shown schematically in Table 9.

TABLE 9
SCHEMATIC OUTLINE FOR REVERSAL EXPERIMENTS

Stimulus Dimensions		Response Assignments for First 10 Trials			Final Assignment
		Control	Reversal	Nonreversal	
1	1	A	B	A	A
1	0	A	B	B	A
0	1	B	A	A	B
0	0	B	A	B	B

To simplify Table 9, the problem is depicted with only two stimulus dimensions, although there were actually four. The stimulus patterns were four-letter consonant clusters. The four pairs of letters were (V, W), (F, G), (X, Y), and (Q, R), each pair constituting a stimulus dimension. The response alternatives were nonsense names, MIB and CEJ. To make a letter-pair relevant, one letter was always present on the MIB cards and the other letter of that pair always appeared on the CEJ cards. One letter of each pair was sampled to construct each of the $2^4 = 16$ letter combinations. The four letters were arranged in an imaginary cross on the card. Two examples are shown below.

Pattern 1

V
Q G
Y

Pattern 2

W
R F
X

The letters from the four pairs were fixed in order but not in location. To equalize the probability that a presented letter would be selected by a subject, the letters were rotated around the four ends of the imaginary cross, yielding 64 different patterns. The order of the letters is the one given in the illustration above. The final relevant dimension for all subjects was the pair (V, W). In Table 9, dimension 1 is (V, W) and dimension 2 might be (F, G).

During the first ten trials, the three groups, totaling 66 subjects, were reinforced according to the stimulus-response assignments shown in the middle three columns of Table 9. The control and reversal groups had the first pair (V, W) as the relevant dimension, but they had opposite response assignments. The nonreversal-shift subjects had a different cue (one of the

other letter pairs) as the relevant dimension during the first 10 trials. If a subject began a criterion run of 16 consecutive correct responses starting on or before trial 10, his response assignments were not changed. But if he made an error on trial 10 or thereafter, before a criterion run of 16 was completed, on the next trial his response assignments were changed to the final assignment given in the right-hand column of Table 9. The relation between the initial and final assignments accounts for the group designations as control, reversal-shift, and nonreversal-shift.

According to all-or-none theory, the error on trial 10 or soon thereafter indicates that the initial response assignments had not been learned. Since an error resets the subject back to zero, all subjects start from the same point (of ignorance) on the final, identical response assignments. Effectively, the theory relies on a single error to give complete information about the subject's state of knowledge. Since all subjects who get to the final problem are in the same state of ignorance, the theory predicts that their behavior on the final problem will be indistinguishable. Thus, for subjects who reach the final problem, the reinforcement contingencies during the initial ten trials (or more, depending on their responses) will have no effect on their final-problem performance.

In contrast to these predictions, incremental theories would seem to predict that during the initial trials subjects were gradually learning incorrect habits and that these would interfere with performance on the final learning task. For example, the Bourne-Restle theory predicts more errors for the reversal group than for the control group. Because of the assumed adaptation of irrelevant cues in this theory, subjects in the nonreversal condition would have a difficult final learning task since the relevant cue in the final problem is one that the subject was learning to ignore during the initial training series. The same predictions are made by Restle's strategy-selection model if it is assumed that the subject eliminates hypotheses that were wrong or irrelevant during initial training (see p. 119, this volume). However, if subjects select hypotheses randomly and with replacement, then the strategy-selection theory predicts no difference among the three groups on the final problem.

Perhaps we should explain why all subjects in the experimental groups were not reversed on trial 10. With such a procedure, an unknown amount of heterogeneity is produced in the groups working on the final problem. Some subjects would have learned by trial 10, and others would not. This may be inferred from whether a subject makes an error on trial 10 or later. Performance on the final problem will differ depending on whether the subject has learned before the shift occurs. If a subject has learned an initial problem, then his performance on a second problem depends markedly on the relation between the response assignments on the two problems. Judging from other results on shifts following learning, nonreversal shifts are more difficult than reversal shifts (Kendler and Kendler, 1962). Thus, if we were to compare on the final problem those subjects who learn the initial problem,

we would expect the controls to be best, the reversal subjects to be next best, and the nonreversal subjects worst on the final problem. However, we are not concerned with these differences in shift rate *after* original learning is completed. By restricting the critical comparison to those subjects who have not learned, as indicated by their error initiating the shift, the theory is thus assured that all subjects who enter into the critical comparison begin the final problem in the same state of knowledge. This restriction avoids contamination of the post-shift results by unknown numbers of subjects who have learned and are expected to differ markedly on the final problem.

Now we turn to the results. Eleven subjects started a criterion run of 16 correct on or before trial 10. This left 18 subjects in each condition for the comparison on the final problem. The distribution of early solvers in groups *C*, *R*, and *NR* was 3, 4, and 4, respectively. These 11 subjects do not enter into the comparisons on the final problem. However, their data will be used in a later analysis.

The critical comparison of the three groups is the comparison of the average total errors before learning the final problem, where responses are counted from the trail immediately following the error-trial that initiated the final problem. Two subjects, one in the control group and one in the nonreversal group, failed to reach criterion within 140 post-shift trails; all other subjects solved before 140 post-shift trials. The results on average errors and average trial of the last error on the final problem are shown in Table 10. Also given in Table 10 are the number of subjects in each condition, the standard deviations, and the ε estimates for the final problem performance of the three groups.

TABLE 10

FIRST REVERSAL EXPERIMENT: MEAN ERRORS AND TRIAL OF LAST ERROR, STANDARD DEVIATIONS, AND ε ESTIMATES FOR THE FINAL PROBLEM

Group	*N*	$\hat{\varepsilon}$	Mean Errors	*SD*	Mean Trial of Last Error	*SD*
Control	18	.052	19.11	19.01	38.33	32.50
Reversal	18	.052	19.11	16.42	39.56	32.25
Nonreversal	18	.055	18.28	19.28	36.94	38.23

The differences between mean errors and mean trail of last error are small with respect to the sampling variability of these scores. The null hypothesis that $\varepsilon_1 = \varepsilon_2\ \varepsilon_3\ \varepsilon_c$ was tested by the likelihood ratio:

$$\lambda = \frac{\varepsilon_c^{N}(1 - \varepsilon_c)^{T-N}}{\Pi_{i=1}^{3}\varepsilon_i^{N_i}(1 - \varepsilon_i)^{T_i - N_i}}.$$

When the three groups were pooled, the common estimate was $\hat{\varepsilon}_c = .053$. After substituting the above estimates for the ε_i and taking $y = -2$ in λ, we

obtained 0.17 as the value of y. With 2 df, $\chi^2 = 0.17$ gives a $P > .90$ for the null hypothesis.

If the initial training did establish differential (wrong) habits, their effect in interfering with the final learning is negligible. Apparently when a subject makes an error he is, in effect, indicating that he knows nothing about the correct solution to the problem. Thus each error is a recurrent event that resets the process to zero. In this regard, when the subjects were questioned after the experiment, only two of them expressed any idea that the correct-response assignments may have changed. Both of these subjects were in the control group, where, in fact, there was no change.

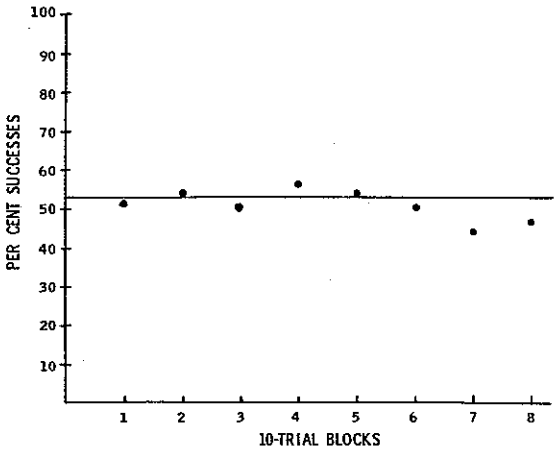


FIG. 6. Stationarity data: percentage of successes prior to the last error, plotted in blocks of ten trials (first reversal experiment).

We now proceed to detailed quantitative analyses of these data. Because previous analyses did not reveal differences between the three groups, the groups were pooled. The two subjects who failed to solve, one each in the control and nonreversal groups, are considered to have solved on their last error (trial 151). Also, the 11 subjects who began their criterion run on trial 10 or before were added to the pool. After equating the number of subjects in the three groups, one extra control subject remained; he made 29 errors on the final problem and his protocol was added to the pool. This pooling results in 66 subjects. These 66 response sequences (from trial 1 forward) were analyzed as if they represented protocols from 66 subjects who were working on the same problem. The aggregation is legitimate because the theory does not distinguish subjects according to their initial S-R assignments.

The forward stationarity curve is shown in Fig. 6. This curve shows the percentage of correct responce in blocks of ten trials, each block containing only those subjects whose final error occurs on some later block. There is

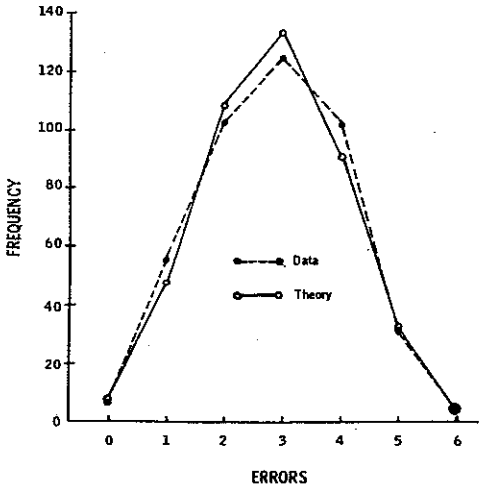


FIG. 7. Theoretical and observed binomial distributions of number of errors in blocks of six trials prior to the last error (first reversal experiment).

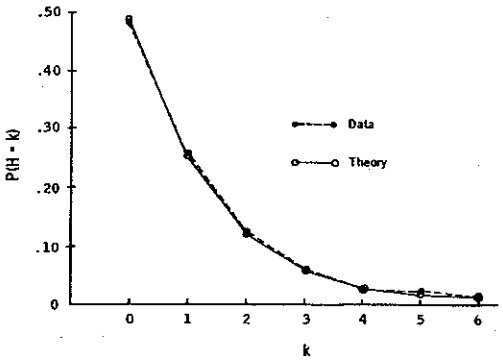


FIG. 8. Distribution of H, the number of successes between adjacent errors (first reversal experiment).

an average of 300 observations in each block. The successive estimates are stationary. The value of χ^2 was 9.73 with 7 df, $P > .20$. The data prior to the last error were also analyzed in Vincentized quartiles, with a total of 662 responses in each quartile. The χ^2 for stationarity was 3.40 (3 df, $P > .30$).

Successive responses prior to the last error were statistically independent. The conditional probability of a success was .534 following a success and .528 following an error ($\chi^2 = .15$, df $= 1$, $P > .50$). This test was based on 2674 observations and had considerable power.

A property of Bernoulli random variables is that the sum of N of them has a binomial distribution with parameters p and N. Figure 7 presents a histogram of the frequency with which exactly 0 to 6 errors occurred in blocks of six trials prior to the last errors, summing over all six-trial blocks for a given subject and then summing over all subjects. Theoretical frequencies are also shown in Fig. 7. The fit is good ($\chi^2 = 4.72$; df $= 6$, $P > .50$).

The distribution of the number of successes between adjacent errors is shown in Fig. 8. The theoretical function is qp^k with $\hat{p} = .491$. A large

number of observations is involved (1,370), and the fit is good. The observed mean is 1.04, with 1.04 predicted; the observed standard deviation is 1.42, with 1.45 predicted.

We now present those predictions that are a function of the learning parameter, ε. Figure 9 shows the cumulative distribution of total errors; the theoretical curve is for $\hat{\varepsilon} = .048$. The maximum discrepancy of .053 occurs at the second block ($N = 66$, $P > .20$), and one cannot reject the fit of the model to these data.

Table 11 compares point predictions of the model with $\hat{p} = .491$, and $\hat{\varepsilon} = .048$. The predictions are reasonably accurate.

The second experiment was a replication of the previous findings regarding reversal shifts, using different stimulus materials (geometric patterns rather than letters) and what was presumed to be an easier problem. An easier problem was chosen to counter the possible objections that the previous results were due to the use of a difficult problem. Our efforts to construct an easier problem were successful; the learning rate on the second problem was about twice as fast as that observed in the prior experiment with letter patterns. If one were to reverse subjects who have not learned by trial 10 on an easier problem, many subjects would be lost for the post-shift comparison because they would learn before trial 10. In order to avoid wasting subjects in this way, our procedure in the second experiment was to reverse a subject if he made an error on trial 5 or later (instead of trial 10, as before). With this procedure, fewer subjects were needed to build up the size of the sample for the final problem. The dilemma confronting us here is a standard optimization problem: we want to use an easier problem, but keep the number of trials before reversal large enough to be convincing, and yet not waste time

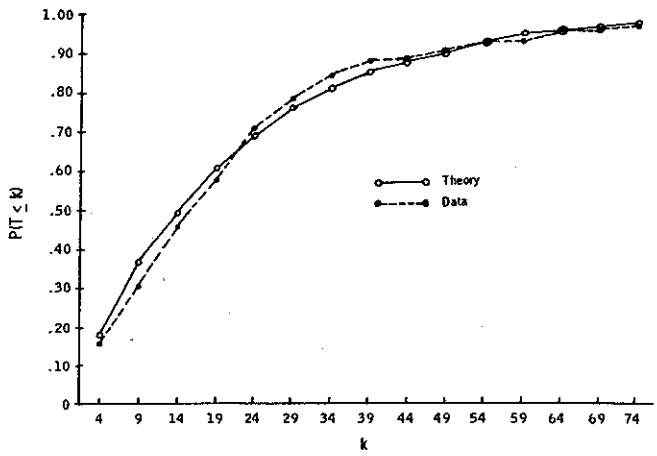


FIG. 9. Cumulative distribution of T, total errors, plotted in class intervals of five errors (first reversal experiment).

TABLE 11

FIRST REVERSAL EXPERIMENT: DETAILED PREDICTIONS

Statistic		Observed	Predicted
Total errors		20.85	20.85
Standard deviation		18.49	20.30
Errors before first success		1.01	.99
Standard deviation		1.39	1.37
Errors before second success		2.17	2.05
Standard deviation		2.29	2.07
Trial of last error		40.94	40.94
Standard deviation		34.59	40.70
Successes between adjacent errors		1.04	1.04
Standard deviation		1.42	1.45
Probability of an error following an error		.46	.48
Alternations of success and failure		22.04	21.06
Runs of errors,	R	11.23	10.72
Runs of 1 error,	r_1	6.33	5.54
	r_2	2.52	2.68
	r_3	1.20	1.30
	r_4	.52	.62
Error pairs,	c_1	9.56	10.11
	c_2	9.67	9.86
	c_3	9.12	9.63
	c_4	9.20	9.40
	c_5	8.75	9.17

and effort with too high a percentage of subjects who learn before they can be put on the final problem. The fact that the trial of learning has a geometric distribution implies that the critical trial for determining reversal must be reasonably small if many subjects are not to be lost because they learn before they can be put onto the final problem.

In the second experiment, the six attributes of the stimuli were shape (square or hexagon), colored area within the figure (upper right–lower left quadrants or upper left–lower right quadrants), number of identical figures (three or four), location of figures (2 in upper left and 1 or 2 in lower right, 2 in upper right and 1 or 2 in lower left), size (large or small), and color of the outline and quadrants of the figures (red or blue). There were $2^6 = 64$ different stimulus cards, representing all combinations of the six two-valued dimensions. The final relevant dimension for all subjects was color, and the response alternatives were the numbers 1 and 2. In Table 9, which outlines the schema of the experiment, the first attribute is color and the second might be shape.

During the first five trials, three groups, totaling 46 subjects, were reinforced according to the stimulus-response assignments shown in the middle

three columns of Table 9. Control and reversal groups had color-relevant but opposite response assignments. The nonreversal subjects had a different relevant cue (selected at random) for the first 5 trials. If a subject began a criterion run of 16 consecutive correct responses on or before trial 5, his response assignments were not changed. If a subject made an error on trial 5 or thereafter before a criterion run of 16 correct was completed, on the next trial his response assignments were changed to the final assignments given in the right-hand column of Table 9.

TABLE 12

SECOND REVERSAL EXPERIMENT: MEAN ERRORS AND TRIAL OF LAST ERROR, STANDARD DEVIATIONS AND ϵ ESTIMATES FOR THE FINAL PROBLEM

Group	*N*		Mean Errors	*SD*	Mean Trial of Last Error	*SD*
Control	10	.078	12.9	8.42	28.6	20.82
Reversal	10	.067	14.9	9.77	29.0	19.71
Nonreversal	10	.071	14.0	14.15	26.9	26.45

Sixteen subjects (of the 46) started a criterion run of 16 correct on or before trial 5. There were 4, 5, and 7 subjects in the *R*, *C*, and *NR* groups, respectively. This left 10 subjects in each condition for comparison on the post-shift trials. All subjects reached criterion in the experiment.

The results on average errors and average trial of the last error on the final problem are shown in Table 12. Although the control group made slightly fewer errors, the differences are small relative to the sampling

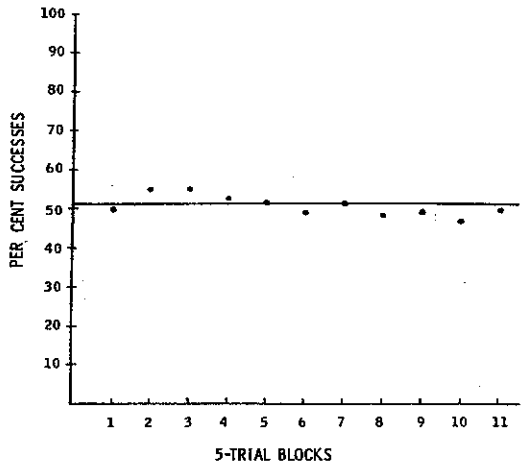


FIG. 10. Stationarity data: percentage of successes prior to the last error, plotted in blocks of five trials (second reversal experiment).

variability of the scores. Moreover, there are negligible differences in the number of trials required to learn the final test problem. The hypothesis that $\varepsilon_1 = \varepsilon_2 = \varepsilon_3 = \hat{\varepsilon}_c$ was evaluated by a likelihood ratio test, using $\varepsilon_c = .072$. The resulting χ^2 was .19 (2 df, $P > .90$). Thus, one may conclude as before that the final-problem performance is not affected by the response assignments reinforced during the initial series.

Protocols of all 46 subjects were pooled for quantitative analyses. The forward stationarity curve (in five-trial blocks) is shown in Fig. 10. An average of 90 observations entered into each block. The hypothesis of stationarity is supported ($\chi^2 = 2.79$, df = 10, $P > .98$). Vincentized quartiles of responses prior to the last error (227 responses per quartile) were also stationary ($\chi^2 = .79$, df = 3, $P > .85$). Successive responses prior to the last error were statistically independent. The conditional probability of a success after a success was .50 and after a failure was .52. The χ^2 based on 932 observations was .57 (1 df, $P > .40$).

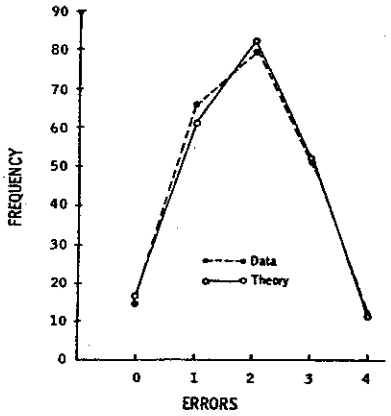


FIG. 11. Theoretical and observed binomial distributions of number of errors in blocks of four trials prior to the last error (second reversal experiment).

The predicted and observed binomial distribution of errors in blocks of four prelearning trials is shown in Fig. 11. The χ^2 for goodness of fit was 1.04 (4 df, $P > .90$). Figure 12 shows the distribution of the number of successes between adjacent errors. The p value is .50, yielding a good fit. The observed mean is .998, with 1.00 predicted; the observed standard deviation is 1.48, with 1.42 predicted. The $\hat{\varepsilon}$ was .087, about twice as large as it was in the previous experiment. This estimate was used to predict the cumulative distribution of total error scores shown in Fig. 13. The maximum discrepancy is .082, occurring on error block 15–17 ($P > .20$). Thus, one cannot reject the model's fit to the total error distribution.

Table 13 presents point predictions for the second reversal experiment. For these predictions, $\hat{p} = .50$ and $\hat{\varepsilon} = .087$. The predictions are very accurate. Especially impressive are the predictions of the variance of total errors and the variance of the trial of the last error.

To summarize, the two reversal experiments tested the prediction that final problem performance would be comparable for control, reversal, and nonreversal subjects who are equalized at the beginning of the final problem by the requirement that they make an error before they start the final problem. Both experiments yielded results supporting this prediction. A second prediction that follows if ε remains constant over trials is that the conditional

distribution of errors made by subjects on the post-shift trials should be the same as the unconditional distribution of errors for all subjects from trial 1. The results are equivocal on this prediction. In the first experiment, all subjects pooled from trial 1 averaged 20.85 errors, whereas the average post-shift errors on the final problem by the three groups was 18.83. In the second experiment, the unconditional mean errors was 11.45, while the conditional mean for post-shift trials was 13.60. Thus, where nearly equal means were expected, in the first experiment a small decrease and in the second experiment a small increase in the conditional mean errors is observed. The difference between the two means averages out to about zero for the two

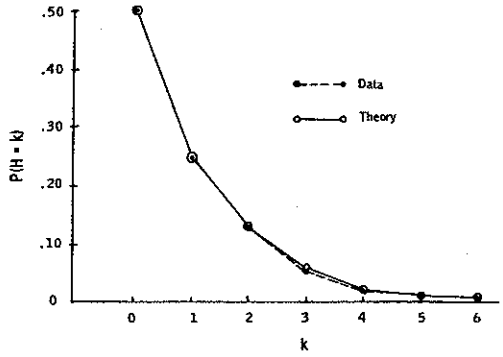


FIG. 12. Distribution of H, the number of successes between adjacent errors (second reversal experiment).

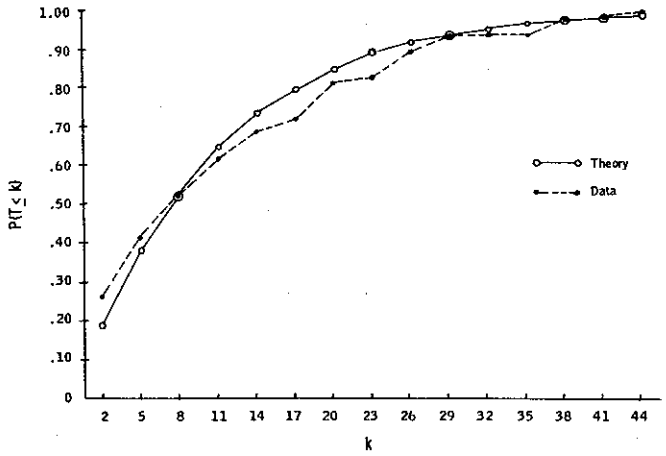


FIG. 13. Cumulative distribution of T, total errors, plotted in class intervals of three errors (second reversal experiment).

TABLE 13

SECOND REVERSAL EXPERIMENT: DETAILED PREDICTIONS

Statistic		Observed	Predicted
Total errors		11.45	11.45
Standard deviation		11.02	10.96
Errors before first success		.91	.92
Standard deviation		1.15	1.27
Errors before second success		1.87	1.78
Standard deviation		1.82	1.73
Trial of last error		22.89	22.90
Standard deviation		22.39	22.40
Probability of an error following an error		.46	.46
Alternations of success and failure		11.87	11.95
Runs of errors,	R	6.17	6.18
Runs of 1 error,	r_1	3.11	3.34
	r_2	1.72	1.55
	r_3	.76	.71
	r_4	.39	.33
Error pairs,	c_1	5.28	5.23
	c_2	4.87	4.95
	c_3	4.70	4.70
	c_4	4.71	4.47
	c_5	4.30	4.26

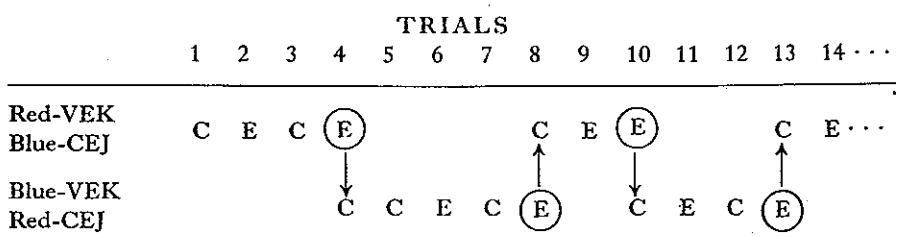
experiments combined. Because the differences between the conditional and unconditional means are small and their rank orders are reversed in the two experiments, no rational explanations of the differences are offered. The small differences in means are presumed to result from sampling variability associated with error scores.

The next study has considerable power in differentiating between theoretical assumptions. The two prior experiments relied upon a single reversal of the S-R assignments to show that an error met the requirements of a recurrent event. The next study carries the reversal procedure to an extreme: on every second error the subject made, the S-R assignments were reversed. Thus, as the subject proceeded, the response assignments were repeatedly changing back and forth.

The stimulus patterns were the colored geometric forms used in the second reversal experiment except that the quadrants that were colored were fixed. The stimuli varied in color (red or blue), shape (square or hexagon), number (three or four figures), size (large or small), and location of figures (2 upper left or 2 upper right). Color was the relevant attribute throughout the experiment. The 16 subjects[2] in the control group learned a color concept,

[2] The subjects for this experiment, students from Foothill Junior College, were obtained through the kind assistance of Dr. Milton Kielsmeier.

half with each *S-R* assignment (VEK or CEJ to red or blue figures), to a criterion of ten consecutive correct responses. For a subject in the experimental group ($N = 17$), the assignment of responses was reversed on every second error. On his second error of a subseries, the subject was told "Correct," in accord with the instantaneous reversal of assignments the experimenter made as soon as the subject responded. The diagram below illustrates the first fourteen trials for a hypothetical subject who begins with the assignments Red-VEK, Blue-CEJ. With respect to this assignment, his first and third responses are correct while his second and fourth responses are errors. His second error, on trial 4, is circled to indicate that immediately following the subject's response on that trial the experimenter said "Correct." At this point the experimenter has shifted to reinforcing the reversed *S-R* assignments given in the bottom row. With respect to these reversed assignments, the fourth, fifth, and seventh responses were correct, and the sixth and eighth responses were wrong. Because the response on trial 8 was the second error of the subseries, the *S-R* assignments were reversed again, back to the original assignments, and the subject was told "Correct" for his response on trial 8. The series of reversals continued in this fashion until the subject met a criterion of ten consecutive correct responses following a reversal. With this procedure, the reason that assignments were reversed every second error is apparent; if assignments were reversed on every single error, the subject would always be told "Correct." In such circumstances, the experimenter has no control over what the subject learns.



The assumption that learning occurs only on error trials predicts that, for the reversal subjects, the number of "called" errors (cases in which the subject's response is disconfirmed) should be the same as the total errors made by the control subjects. This prediction follows from two parts of the theory: (*a*) that learning occurs on a single trial, and (*b*) that the only effective learning events in this situation are "called" errors. The alternative assumption that learning may occur on either correct or incorrect trials predicts (*a*) that the average trial of the last error will be the same for the two conditions, and (*b*) that the number of errors (called and uncalled combined) for the experimental subjects will be equal to the number of called errors for the control subjects.

The predictions from incremental theory differ from the predictions

above. One interpretation of incremental theory is that subjects in the reversal condition would never reach criterion. However, there are special circumstances under which an incremental theory would expect some subjects to reach criterion. Such learning would depend upon a fortuitous trapping effect. If a response is called correct, then that habitual mode of responding is reinforced. Hence it is more likely than before that another response, controlled by the same attributes, will occur and will be called correct, and the habit will be strengthened further. The positive feedback in the system is obvious; if a correct response occurs, the probability of another correct response is increased, and this positive feedback could draw the subject into a criterion run of 10 consecutive correct responses. We will not go into further details concerning this type of explanation. It is clear, however, that the rate at which the reversal subjects reach criterion should be slower than the rate at which the control subjects reach criterion.

All subjects but two (one in each group) met the learning criterion within the experimental time limit. The two nonsolvers are excluded from further analyses. The 16 subjects in the reversal group averaged 7.00 reversal shifts before meeting the learning criterion. The important fact is that the average number of "called" errors was about the same for both groups: 7.81 "called" errors for the reversal subjects and 8.00 for the 15 control subjects. The standard deviation of total errors for the control subjects was 8.22. Thus, the difference of .19 in the average number of called errors is small with respect to the variance of this measure.

Two subjects in each group learned after making only one error. Thus, two subjects in the reversal condition were never reversed because they learned their initial response assignments. Disregarding subjects with only one error, the mean number of called errors was 8.79 for the reversal subjects and 9.08 for the control subjects. The mean number of reversals was 8.00. Again the differences are negligible. Thus, on the question of primary interest, the results favor the all-or-none theory with learning on errors.

Using the theory, one also may predict the mean trial of the last error for the reversal group once the mean total errors for the control subjects is known. The parameter p is assumed to be .50. Let T_r be the mean number of called errors for the reversal subjects, and let r be the mean number of reversals before they learn. Then the average trial of the last error for the reversal subjects will be

$$n_r' = T_r + r + 1 + 2(T_r - 1). \tag{55}$$

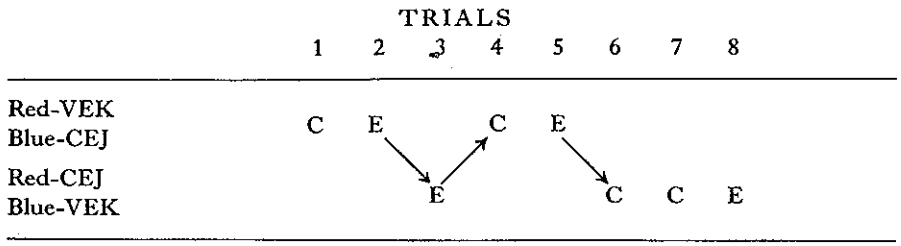
The terms T_r and r in Eq. (55) are clearly needed. The additional terms $1 + 2(T_r - 1)$ are the expected number of correct responses (excluding calls on reversal trials) for a subject who makes T_r errors in the reversal group. The T_r called errors partition the successes as follows: there is an average of one success before the first error, and subsequently an average of two successes between each of the $T_r - 1$ "called" errors.

The right-hand side of Eq. (55) can be written as a simple function of T_c, the mean number of errors in the control group. First, by the error-learning assumption, "called" errors should be the same for both groups, so $T_c = T_r$. Secondly, the number of reversals, r, is related to T_r in a direct manner. If a subject makes T_r called errors, then his number of reversals will be $T_r - 1$, assuming that he makes one more error after his rth and final reversal. If we substitute into Eq. (55) the identities $r = T_r - 1$ and $T_r = T_c$, the relation between T_c and the average trial of the last error for the reversal subjects is

$$n_r' = 4T_c - 2. \tag{56}$$

The observed T_c for the 15 solvers in the control group was 8.00. Substituting this value into Eq. (56), we obtain $n_r' = 30.00$. The observed trial of last error for the 16 solvers in the reversal group was 28.81, so the free prediction across groups is close.

Finally we report results of a pilot study relating to the prior study. The results are relevant to interpreting the subject's behavior on the trial when he learns. In the prior study, the *S-R* assignments were reversed on every second error, although the subject was told "Correct" on that trial. With that procedure, it is not feasible to reverse the assignments on every single error since the subject would always be told "Correct." This procedure was modified so that the subject was told when he made an error and the *S-R* assignments were reversed on the next trial following each error. The diagram below illustrates the procedure for the first eight trials of a hypothetical subject in this study. Using the same stimulus materials as in the previous experiment, with color relevant, the response assignments were changed on the trial following each error. Eleven college students were trained on this problem, either until they gave 15 consecutive correct responses or until they made their first error after trial 100, whichever occurred first.



Although this procedure represents only a minor modification of that used in the prior reversal study, analysis shows that the modifications are critical and that learning would be expected to be extremely slow. The difficulty stems from the fact that a correct hypothesis formulated by the subject following his error on trial n will be wrong when he tests it on trial $n + 1$.

Consider the cognitive events that occur immediately after the subject

receives an error-feedback signal. The subject resamples the set of cues and selects the relevant attribute with probability r, the relative weight of that cue. Thus, with probability r the subject will next try an hypothesis based on the relevant cue, say, color. If the dimensions are binary, then there are two symmetric color hypotheses: "red is VEK and blue is CEJ," and its converse, "red is CEJ and blue is VEK." Suppose that the pattern on trial n is red, that the subject says "CEJ," and that the experimenter says "No, VEK is correct." If this information is perfectly retained when the subject constructs his color hypothesis, then red-VEK, blue-CEJ is *consistent* with the information, whereas the converse hypothesis is inconsistent. Assume that with probability $1 - f$ the hypothesis is consistent with the information received on the trial, and that with probability f the hypothesis is inconsistent.

Why should a subject try a hypothesis that is inconsistent with the information he has just received? He may do so for several reasons. One likely explanation is that he forgets part of the trial-information before he has completed his stimulus scanning and selected a cue to test. If the correct answer is presented and removed quickly, then the subject may forget what the answer was. If the stimulus pattern is removed before the answer is given, then the subject may forget the specific values of the attributes on that pattern. This forgetting occurs over very brief intervals (2 or 3 seconds) after the information event and before the subject reconstructs his new hypothesis. The occurrence of forgetting over short-time intervals is a well-documented fact (Peterson and Peterson, 1959).

We identify f with the probability that one or another of these items of information is forgotten by the time the subject constructs his new hypothesis. In general, f will be small and is expected to vary with experimental conditions. For example, if the stimulus pattern is removed t seconds before a brief feedback signal is given, f should increase exponentially with t. In the pilot experiment, the stimulus card was left in view for three seconds after the experimenter gave the correct answer. Hence, in this context, f would be interpreted as the probability that the subject forgets the answer (spoken by the experimenter) before his new hypothesis is constructed.

Assume that the subject constructs his hypothesis consistent with the trial-information as he recalls it. Thus, with probability $1 - f$ his recall is veridical, so his hypothesis is consistent; with probability f the trial-information is incorrectly recalled, so his hypothesis is inconsistent with the actual information given on that trial. In the conventional learning task, where the S-R assignments remain fixed, only consistent relevant hypotheses are correct. Thus, the average probability of solving after an error is $\varepsilon_1 = r(1 - f)$. When the S-R assignments are reversed after every error, only inconsistent relevant hypotheses solve the problem. Hence, the average probability of solving after an error is $\varepsilon_2 = rf$. Since the amount of short-term forgetting is expected to be small, learning should be extremely slow with the every-error-reversal procedure. By comparing this learning rate with that of control

subjects trained with fixed assignments, an estimate of f, the amount of short term, intratrial forgetting, can be obtained.

In the pilot study with the every-error-reversal procedure, 8 of the 11 subjects did not solve the problem within 100 trials. The mean trials to criterion for the 3 solvers was 47.7. The mean percentage of successes before the last error was .499. Nearly all subjects expressed considerable frustration and confusion about what was going on, and several of the nonsolvers felt that there was no solution to the problem.

By modifying the likelihood function to take account of the number of solvers, N_s, the following maximum likelihood estimate of ε may be derived:

$$\varepsilon = \frac{N_s}{T - (N - N_s)}. \tag{57}$$

The observed statistics were $T = 478$, $N = 11$, and $N_s = 3$. Thus,

$$\hat{\varepsilon}_1 = rf = .00638.$$

A control estimate of ε when the S-R assignments remain fixed was obtained from the 16 control subjects in the previous experiment. For that group, $T = 153$, $N = 16$, and $N_s = 15$, and hence $\hat{\varepsilon}_2 = r(1 - f) = .0988$. From the values of $\hat{\varepsilon}_1$ and $\hat{\varepsilon}_2$ an estimate of f is obtained:

$$\frac{\varepsilon_1}{\varepsilon_2} = \frac{f}{1 - f} = \frac{.00638}{.0988} = .0646; \tag{58}$$

therefore, $\hat{f} = .06$ and $1 - \hat{f} = .94$. The estimate $1 - \hat{f} = .94$ is interpreted as the probability that the subject retains the correct information over the brief time before he constructs his hypothesis. The estimate is sensible and agrees quantitatively with the retention probabilities for 2- to 3-second intervals reported by Peterson and Peterson (1959).

8. Discussion

The theory proposed for concept-identification experiments views the main behavioral process as the selection of the relevant stimulus attributes. This selection process has been described in terms of the subject's attending or observing responses. The word "response" is used in a hypothetical sense, since one usually cannot specify the physical topography of such responses. In a few cases, the physical topography might be specified in terms of receptor-orienting acts, e.g., turning the eyes to look at the relevant letter in the leftmost position on the card. But in most cases, one cannot specify the topography of the attending responses learned, except in the trivial sense that subjects learn to orient towards the place where the stimuli are presented in the experimental room. The selective responses to which we refer have been described variously as a perceptual set to react to a particular cue, an implicit encoding response, or a mediating response. A number of theorists have discussed the bearing of such selection mechanisms upon the results of concept-learning experiments (e.g., Garner, 1962; Hunt, 1962; Kendler and Kendler,

1962; Lawrence, 1963; Osgood, 1953; Shepard, Hovland, and Jenkins, 1961).

The discussion below interprets our results in terms of perceptual encoding, using primarily the terms introduced in a paper by Lawrence (1963). An encoding process is a method by which a subject constructs some internal representation of a stimulus pattern. From the viewpoint of a human subject, the encoding process can be characterized by the kind of questions he asks about the stimulus pattern. Thus, "How many objects are there?" would be an example of an elementary encoding process, and the stimulus-as-coded (s-a-c) might be the implicit description "four." The s-a-c is the functional unit that enters into response selection, and it is the unit that becomes associated with overt instrumental responses. Aspects of the stimulus display that are not encoded or employed in the representation on a particular trial are not modified in their associative connections to the reinforced response. Stimulus aspects that are not noticed are unaffected by the reinforcement contingencies.

Nearly all stimuli can be described, judged, or classified in a multitude of ways, and each of these ways indicates the use of a different encoding response. The coding response usually depends upon factors other than the proximal stimulus. In most instances, the coding process is determined by prior learning. With human subjects, the coding response can be modified instantaneously by instructions, e.g., a subject is instructed first to judge a series of tones by their pitch, and then by their loudness. According to this approach, concept-identification learning involves the acquisition of two sets of habits. Through a trial-and-error process, the subject selects a relevant coding response, and then learns associations between the resulting s-a-c's and the overt classificatory responses. The overt responses become conditioned to the relevant s-a-c's; the implicit coding response is conditioned, presumably, to those stable and unchanging background or contextual stimuli associated with the experimental situation.

For the familiar attributes used in our experiments (colors, shapes, etc.), we suppose that the subject has available a coding response for each attribute. The coding response in a college-student population would be indexed, presumably, by the ability to verbalize the name of the attribute. Thus, if a subject inspected all the patterns in the series, presumably he could name or describe the attributes that varied over the series. The order in which the subject gives these names or descriptions would be determined by the saliency of the attributes. This notion suggests a series of experiments in which the learning rate when cue x is relevant is correlated with the empirically determined rank order with which x is described as an attribute of the stimuli. For a given population of subjects, such empirical norms could be obtained and used to predict relative learning rates.

In the preceding discussion it is assumed that a subject gives descriptive names of attributes according to a given hierarchy. The presumption is that

during learning the subject tries out a description (name, coding response) on each trial, and that the relative frequency of usage of the various coding responses varies directly with their position of dominance in the subject's hierarchy of coding responses. The relative weight for cue i, $w_i/\Sigma w_j$, may be interpreted as the average probability that before learning the subject selects the mediating response that encodes the stimulus pattern by the name of cue i. When learning occurs with cue i relevant, its coding response is placed at the top of the hierarchy and it occurs on every trial. Stated in this manner, the proposed stimulus-selection process in concept-identification learning resembles the spew hypothesis proposed in other contexts by Bousfield and Sedgewick (1944) and by Underwood and Schulz (1960). Bousfield and his colleagues have employed this notion in describing the order of emission of responses in free recall; Underwood and Schulz applied the hypothesis to the response-recall process in paired-associate learning. The research potential of the hypothesis for concept learning has not yet been explored sufficiently.

The hierarchy of coding responses can be manipulated by stimulus characteristics (e.g., some attributes may be emphasized), by verbal instructions, or by past learning. The role of past learning is shown in a series of transfer studies comparing reversal and nonreversal shifts in concept learning. Following learning to a criterion with dimension A relevant, with value A_1 paired with response R_x and value A_2 paired with response R_y, subjects in the reversal condition then learn the new pairings $A_1 - R_y$ and $A_2 - R_x$. Subjects undergoing a nonreversal-shift must learn to attend to a different, previously irrelevant cue, say cue B, and must learn the pairings $B_1 - R_x$ and $B_2 - R_y$. The comparative difficulty of reversal and nonreversal shifts depends upon the apparent availability of (verbal) mediating responses in the subjects being studied (Kendler and Kendler, 1962). College students perform the reversal-shift faster, rats do the nonreversal-shift faster, and kindergarten children learn about equally fast on the two shifts. Presumably, college students use verbal mediating responses in solving the original problem. In a reversal-shift, they learn to associate the opposite responses with the values of the relevant s-a-c's. In a nonreversal-shift, the subject must learn to attend to a different cue and thus to rearrange the hierarchy of coding responses established in learning the original problem.

Other evidence for a mediating process comes from studies of intradimensional shifts. In an intradimensional shift, the specific values of all dimensions are changed between the first and second problems, but the same attribute (e.g., color) is relevant in both problems. Thus, in the first problem, color might be red or blue; in the second problem, color might be yellow or brown. In an intradimensional shift, the subject need only learn two paired-associates (involving the new values of the relevant attribute), and the mediating response hierarchy need not be rearranged as in an extradimensional shift (new values and a different relevant attribute). Results by Zeaman *et al.* (1961)

and by Bower and Gadberry (unpublished) show the superiority of subjects performing on the intradimensional shift.

The attending or mediating responses often are not observed and recorded; they are inferred, hypothetical constructs. However, experimental conditions can be arranged so that attending responses are recorded, observable behavior. Employing a procedure analogous to the procedure Wyckoff (1952) used with pigeons, Bower (unpublished) modified the conventional concept-identification experiment to record both observing responses and classificatory responses. Three pairs of consonant letters (X, Y), (Q, V), (M, F) were used as cues. Random sequences of the eight possible three-letter combinations were constructed. Instead of viewing all three letters of each pattern, the subject could choose to see only one of the three letters by asking to see a red, blue, or yellow information card. After the subject was shown the letter on the requested colored card, he then classified the "hidden" pattern as a 1 or a 2. The blue card carried the relevant information (Y or K). If X were scheduled to appear on the blue card, then the subject should say "1"; if K were scheduled, the subject should say "2." The red and yellow cards contained the two pairs of irrelevant letters, (Q, V) on the red card, (M, F) on the yellow card. When the subject chose to see the letter on one of these cards, the best he could do was 50 per cent correct. In solving the problem, the subjects, of course, learned to ask for the relevant blue card. This technique makes the observing responses an explicit part of the problem-solving process so that they can be recorded and studied in their own right. The college students trained with this procedure did not find it a difficult task. After the subjects learned to ask for the blue information card and learned the correct associations to the letters on the blue card, all new letters were substituted without any warning or break in the experimental routine. Half of the subjects learned the new letter problem with the blue card still carrying the relevant letters (an intradimensional shift); the remaining subjects had to change over to choosing the yellow information card since it bore the relevant letters for classification in the second problem (an extradimensional shift). The difference between the two conditions in learning the second problem was enormous: the intradimensional-shift subjects averaged .50 classification errors before learning; the extradimensional-shift subjects averaged 30.23 classification errors before learning the second problem. The basis for the difference between conditions is obvious. Although these results are not surprising, they clearly show the magnitude of effects that can be produced by manipulating through prior learning the hierarchy of observing responses. This technique for producing overt analogs of mediating responses may prove to be a valuable tool for future research. For example, it may yield useful data on four-category concept problems in which the classification of a pattern depends upon the conjunction of values from two attributes (e.g., color and shape). In this situation, the subject could ask for information on two attributes during each trial.

In concluding this commentary, we should point out that all of our experiments have employed the simplest concept problems. We have employed stimulus materials with easily identifiable properties, and have studied two-category (Type I) concepts defined in terms of common perceptual attributes of the patterns assigned to a given class. We are aware that much of the current experimental literature on concept learning deals with more complex problems, involving, for example, conjunctions, disjunctions, or relations between simple concepts, and that many important facts are known about factors influencing the learning of such complex concepts. Hunt (1962) has reviewed a good deal of this material in his recent book. We have nothing original to add to that literature since we have not yet investigated complex concepts. One of the difficult tasks before us is to extend and modify the simple theory proposed here to account for the way in which subjects learn compound concepts. If simple concepts are learned in all-or-none fashion, it is clear that learning to conjoin two concepts will appear, phenotypically, not to be an all-or-nothing process. Multiple-stage models, of the type proposed by Restle (this volume, p. 124), would seem to offer the more promising leads for extending the simple model to handle the more complex cases.

9. Summary and conclusions

We have presented a theory and model for two-category (Type I) concept-identification learning. Learning in such situations is conceived primarily in terms of a selection process whereby the subject identifies the attribute relevant to the classification. The formal model of this process is a two-state Markov chain. The two states refer to the values p and 1 of the probability of correct training. During the course of the experiment some random (learning) event happens that moves the subject from the initial p-state to the terminal 1-state. The major evidence for the assumption that response probabilities can be represented by only two values came from analyses of responses prior to the last error. Such analyses showed that the sequence of responses prior to the last error could be represented as an independent Bernoulli-trials process with constant probability p of a correct response.

An hypothesis proposed related the learning rate to the average probability that the subject samples or attends to the relevant attribute. This hypothesis was confirmed by predicting results on cue additivity. Some evidence was adduced to support the view that in our problems, the subject's performance is controlled by the relevant cue and is not dependent upon the continued presence of irrelevant cues.

The hypothesis that subjects modify their performance only after making an error was tested against the alternative that subjects may learn equally well on correct- or incorrect-response trials. In favor of the first alternative, evidence was reviewed on the effect of partially valid irrelevant cues. To this evidence was added an experiment comparing two- and four-response alternatives. The number of response alternatives affects the probability of a

correct response prior to the last error, but has no effect on the average errors before learning. These results indicate that the effective learning event occurs on error trials; the subject resamples from the population of cues for testing only when a particular cue has led to a mistake.

The final set of experiments provided various tests of the one-trial learning assumption in conjunction with the hypothesis that the probability of learning following an error is constant and independent of the number of preceding trials or preceding errors. The first two of these studies employed the procedure of changing the response assignments when the subject made an error following some critical trial. From the error, it is inferred that the subject has not yet solved and is still in the same state in which he started. Accordingly, the theory predicts that the final-problem performance will be the same for subjects who were trained initially with differing S-R assignments before they made the error initiating the final problem. The evidence from both experiments confirmed the prediction: performance on the final problem was essentially the same for subjects in the control, reversal-shift and nonreversal-shift conditions. The third experiment reversed the S-R assignments after every second error that the subject made. The mean number of "called" errors before learning was the same for the reversal subjects as for control subjects, who received no reversals in S-R assignments throughout the course of the experiment. The result supports the notion that in Type I problems, learning is all-or-none, errors are recurrent events, and effective learning occurs only on error trials.

Finally, the selection process was interpreted in terms of observing and verbal mediating responses.

(*References follow Appendix, p. 92.*)

Appendix

The theory proposed for concept identification conceives of two successive processes: a stimulus-selection phase whereby the subject comes to attend principally to the relevant attribute, and a paired-associate phase during which the subject learns associations between the values of the relevant dimension and the classificatory responses. With binary stimulus and response dimensions, the terminal paired-associate phase is brief and the model for stimulus selection gives an adequate description of the data. When the number of stimulus and response values is increased, the terminal paired-associate phase begins to constitute a substantial portion of the later trials. If there are $N > 2$ values of the relevant dimension, each to be assigned one of N classificatory responses, then following the stimulus-selection phase it is likely that the subject will know some of the correct associates but not others. Accordingly, the average probability of a correct response would increase above the initial chance level—that is, with more than binary stimulus and response dimensions, the theory does not predict stationary backward learning

curves at the chance level. We have some evidence for this absence of stationarity. For example, in the four-response experiment, the Vincentized quartiles of trials before the last error gave correct response probabilities of .242, .337, .337, and .457 for the first to last quartile. These values gave a borderline $\chi^2 = 8.09$ (df $= 3$, $.05 > P > .02$).

In this appendix, a general formulation is given that takes into explicit account the terminal paired-associate phase. The model is presented for the general case in which there are N values of the relevant dimension, each occurring with probability $1/N$ in the random series, and in which there are N classificatory responses assigned one-to-one to the N values of the relevant dimension. By specialization of the general model, for $N = 2$ the simple selection model employed throughout the paper is obtained. The response sequences to which the general model applies are the series of correct and incorrect responses, not differentiated according to the stimulus presented.

The general theory essentially attaches the stimulus-selection process onto the front end of the N-element pattern model developed by Estes (1959a). State 0 is defined as the state of the subject during the stimulus-selection phase, before he has selected the relevant attribute for testing. After the subject begins attending to the relevant dimension, there are N equivalence classes of patterns to be conditioned. Each value of the relevant dimension defines, under the subject's encoding operation, a class of equivalent patterns, consisting of N^i members when there are i irrelevant dimensions with N values each. Each equivalence class (or, alternately, each value of the relevant attribute) is identified with a stimulus element. Each element may or may not be connected to its correct response. We identify states j, $1 \leq j \leq N$, of the Markov learning process with the subject's state when he is attending to the relevant attribute and when exactly j of the N values are connected to their correct responses. Let $C_{j,n}$ denote the event of the subject's being in state j at the beginning of trial n. The one-trial, state-to-state transition probabilities, $P(C_{j,n+1} \mid C_{i,n})$, are abbreviated as p_{ij}.

In line with other developments (Bower, 1961), it is assumed that the paired-associate learning between a value and its reinforced response occurs in all-or-none fashion. When the subject is attending to the relevant dimension and an unconditioned value is presented and reinforced, then with probability θ that association is learned. If the N values of the relevant attributes are presented randomly with equal relative frequencies, then the transition probabilities are as follows for $i > 1$:

$$p_{ij} = \begin{cases} \left(1 - \frac{i}{N}\right)\theta & \text{if } j = i + 1, \\ 1 - \left(1 - \frac{i}{N}\right)\theta & \text{if } j = i, \\ 0 & \text{otherwise.} \end{cases} \tag{1a}$$

The factor $1 - i/N$ is the probability that an unconditioned pattern is presented on a trial when i elements are conditioned. Thus $(1 - i/N)\theta$ gives the probability that an unconditioned element is presented and becomes connected; hence, $i + 1$ elements are conditioned after the trial. The transition probability for leaving state 0, where selection of the relevant attribute occurs, is that used in the earlier discussions:

$$p_{0i} = \begin{cases} q\varepsilon & \text{for } i = 1, \\ 1 - q\varepsilon & \text{for } i = 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2a}$$

In other words, stimulus selection occurs with probability ε when the subject makes a chance error with probability q. Elsewhere we have identified ε as $r\theta$, where r is the probability of selecting the relevant attribute and θ is the probability of conditioning the presented value to the reinforced response.

In the matrix below, the assumptions are illustrated for the case of four relevant values and responses. The subject begins in state 0 and eventually ends in the absorbing state 4.

(3a)

State on trial n	State on trial $n+1$: 4	3	2	1	0	P (Correct)
4	1	0	0	0	0	1
3	$\frac{\theta}{4}$	$1 - \frac{\theta}{4}$	0	0	0	1
2	0	$\frac{2\theta}{4}$	$1 - \frac{2\theta}{4}$	0	0	$\frac{3}{4}$
1	0	0	$\frac{3\theta}{4}$	$1 - \frac{3\theta}{4}$	0	$\frac{2}{4}$
0	0	0	0	$q\varepsilon$	$1 - q\varepsilon$	$\frac{1}{4}$

Some axioms relating the state of conditioning to the probability of a correct response are required. If a sampled element is conditioned, then the correct response occurs. However, when the sampled element is not conditioned, then several alternative assumptions can be employed. The appropriate assumption will depend upon procedural and subject variables. The specific assumption tried is that when the subject is confronted with unconditioned patterns, he guesses, using only those responses that he has not yet learned. This assumption seems plausible for intelligent subjects who know that there is a one-to-one correspondence between values of the relevant attribute and the classificatory responses. Thus, when i elements (and responses) are conditioned, the probability of guessing correctly on an

unconditioned element will be $1/N - i$. Let $x_n = 0$ be the event of a correct response on trial n; then the response axioms can be stated as follows:

$$P(x_n = 0 \mid C_{j,n}) = \begin{cases} 1 & \text{if } j = N, \\ \dfrac{j+1}{N} & \text{if } 0 \leq j \leq N-1. \end{cases} \tag{4a}$$

The factor $j + 1/N$ arises directly from our assumption about restricted guessing, viz.,

$$P(x_n = 0 \mid C_{j,n}) = \frac{j}{N} + \left(1 - \frac{j}{N}\right)\frac{1}{N-j} = \frac{j+1}{N}.$$

These response probabilities have been listed opposite the corresponding states of the matrix in Eq. (3a). Response probability reaches unity when state $N - 1$ is entered. For $N = 2$, errors can occur only in state 0, and a stationary backward learning curve is expected. Thus, for $N = 2$ the general model in Eq. (3a) reduces to the one used throughout this paper.

A few results are easily derived for the general Markov chain. Useful results are the distribution of trials and errors in each of the transient error states, 0 to $N - 2$. Define n_i as the number of trials in transient state i. Then the probability distribution of n_i is

$$P(n_i = k) = \begin{cases} q\varepsilon(1 - q\varepsilon)^{k-1} & \text{for } i = 0, \\ \dfrac{(N-i)\theta}{N}\left[1 - \dfrac{(N-i)\theta}{N}\right]^{k-1} & \text{for } 1 \leq i \leq N-1, \end{cases} \tag{5a}$$

having means

$$E(n_0) = \frac{1}{q\varepsilon}, \qquad E(n_i) = \frac{N}{(N-i)\theta}. \tag{6a}$$

Define t_i as the number of errors made in transient state i. The distribution of t_0 is $\varepsilon(1 - \varepsilon)^{k-1}$, as was noted before. For $i > 0$, the distribution of t_i is

$$P\{t_i = k\} = \begin{cases} a_i b_i & \text{for } k = 0, \\ (1 - a_i b_i)\, b_i (1 - b_i)^{k-1} & \text{for } k \geq 1, \end{cases} \tag{7a}$$

where $a_i = 1/N - i$ and $b_i = \theta/[1 - a_i(1 - \theta)]$. The mean value of t_i is

$$E(t_i) = \frac{1 - a_i b_i}{b_i} = \frac{1 - 1/N - i}{\theta} \qquad (1 \leq i \leq N-1). \tag{8a}$$

Equation (7a) is identical to the distribution of errors per item for the simple one-element model applied to paired-associate data, with $i = 0$ so that guessing is unrestricted (Bower, 1961).

Information about total errors over the entire course of learning may be

obtained by summing the t_i over the transient error states. Define total errors, T, as

$$T = t_0 + t_1 + \cdots + t_{N-2}. \tag{9a}$$

The distribution of T will be the convolution of the t_i distributions. The mean and variance of T are the sums of the means and variances of the t_i since the t_i are independent random variables:

$$E(T) = \frac{1}{\varepsilon} + \frac{1}{\theta}\sum_{i=1}^{N-2}\left(1 - \frac{1}{N-i}\right),$$

$$\operatorname{var}(T) = \frac{1-\varepsilon}{\varepsilon^2} + \frac{1}{\theta}\sum_{i=1}^{N-2}\left(1 - \frac{1}{N-i}\right) + \frac{(1-2\theta)}{\theta^2}\sum_{i=1}^{N-2}\left(1 - \frac{1}{N-i}\right)^2. \tag{10a}$$

The above expressions apply to the case of a one-to-one correspondence of relevant stimulus values and responses. To handle our case with four stimulus values but only two responses, the response rules for unconditioned patterns are modified. In the case of a many-one assignment between relevant values and responses, restricted guessing seems to be an implausible assumption. Rather, it is assumed that to unconditioned patterns, the subject guesses among the r alternatives with equal probabilities. Thus, for $N > r$, the response rule is

$$P(x_n = 0 \mid C_{j,n}) = \frac{j}{N} + \left(1 - \frac{j}{N}\right)\frac{1}{r}.$$

Using these assumptions, the expected total errors for our four-response [from Eqs. (10a)] and two-response subjects will be

$$E(T_2) = \frac{1}{\varepsilon} + \frac{3}{2\theta},$$

$$E(T_4) = \frac{1}{\varepsilon} + \frac{7}{6\theta}. \tag{11a}$$

The observed average errors for the two groups were $T_2 = 13.36$ and $T_4 = 12.41$. From Eqs. (11a), the difference between the expectations of T_2 and T_4 is

$$E(T_2) - E(T_4) = \frac{1}{3\theta} \mathrel{\hat{=}} .95. \tag{12a}$$

From this equation, the estimate $\hat{\theta} = .351$ is obtained. If we solve for ε using the observed T_4 and θ estimate, the estimate obtained is $\hat{\varepsilon} = .110$.

There are several points to be noted about these estimates of θ and ε. First, the magnitude of θ is well within the range of θ values we have obtained with short lists of paired-associate items. Second, the ε estimate is just a shade less than $\frac{1}{3}\theta$. Since there were three stimulus dimensions to the problem (color, shape, and angle of line through figure), if one assumes equal weights for all cues, then the proportion of relevant cues is $\frac{1}{3}$. Thus, the hypothesis

that $\varepsilon = r\theta$ predicts an ε of .117, compared with the observed estimate of .110. Third, if we use the values $\varepsilon = .110$ and $\theta = .351$, we can predict the standard deviation of total errors using the formula in Eqs. (10a) for the four-response group. The predicted $\sigma(T_4)$ is 8.90 with the observed $\sigma(T_4)$ at 9.74. This prediction is improved over that obtained from the simple one-stage selection model [predicted $\sigma(T_4) = 11.91$]. Fourth, since the initial p-value for the four-response subjects is $\frac{1}{4}$, the average errors before the first success predicted by the general model will be slightly under 3.00 (observed $J_0 = 3.18$). This prediction is improved over that of the simple selection model (predicted $J_0 = 1.79$).

REFERENCES

BINDER, A. M., and FELDMAN, S. E. The effects of experimentally controlled experience upon recognition responses. *Psychol. Monogr.*, 1960, **74**, No. 9 (Whole No. 496).

BLUM, R. A., and BLUM, JOSEPHINE S. Factual issues in the "continuity controversy." *Psychol. Rev.*, 1949, **56**, 33–50.

BOURNE, L. E., Jr. Effects of delay of information feedback and task complexity on the identification of concepts. *J. exp. Psychol.*, 1957, **54**, 201–207.

BOURNE, L. E., Jr., and HAYGOOD, R. C. The role of stimulus redundancy in concept identification. *J. exp. Psychol.*, 1959, **58**, 232–238.

BOURNE, L. E., Jr., and RESTLE, F. Mathematical theory of concept identification. *Psychol. Rev.*, 1959, **66**, 278–296.

BOUSFIELD, W. A., and SEDGEWICK, C. H. W. Analysis of sequences of restricted associative responses. *J. gen. Psychol.*, 1944, **30**, 149–165.

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

BOWER, G. H. An association model for response and training variables in paired-associate learning. *Psychol. Rev.*, 1962, **69**, 34–53. (a)

BOWER, G. H. Some experiments related to a learning model. Paper read at Western Psychol. Ass., San Francisco, April, 1962. (b)

BRUNER, J. S., GOODNOW, J. J., and AUSTIN, A. *A study of thinking*. New York: Wiley, 1956.

BUSH, R. R. Sequential properties of linear models. In R. R. Bush and W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford, Calif.: Stanford Univer. Press, 1959. Pp. 215–227.

EHRENFREUND, D. An experimental test of the continuity theory of discrimination learning with pattern vision. *J. comp. physiol. Psychol.*, 1948, **41**, 408–422.

ESTES, W. K. Component and pattern models with Markovian interpretations. In R. R. Bush, and W. K. Estes, (Eds.) *Studies in mathematical learning theory*. Stanford, Calif.: Stanford Univer. Press, 1959. Pp. 9–52. (a)

ESTES, W. K. Statistical models for recall and recognition of stimulus patterns by human observers. In Marshall C. Yovits (Ed.), *Self-organising systems: proceedings*. London: Pergamon, 1959. Pp. 51–62. (b)

GARNER, W. R. *Uncertainty and structure as psychological concepts.* New York: Wiley, 1962.

GORMEZANO, I., and GRANT, D. A. Progressive ambiguity in the attainment of concepts on the Wisconsin Card Sorting Test. *J. exp. Psychol.*, 1958, **55**, 621–627.

HOVLAND, C. I. A set of flower designs for experiments in concept formation. *Amer. J. Psychol.*, 1953, **66**, 140–142.

HUGHES, C. L., and NORTH, A. J. Effect of introducing a partial correlation between a critical cue and a previously irrelevant cue. *J. comp. physiol. Psychol.*, 1959, **52**, 126–128.

HUNT, E. B. *Concept learning.* New York: Wiley, 1962.

KENDLER, H. H., and KENDLER, TRACY S. Vertical and horizontal processes in problem solving. *Psychol. Rev.*, 1962, **69**, 1–16.

KRECHEVSKY, I. A study of the continuity of the problem-solving process. *Psychol. Rev.*, 1938, **45**, 107–133.

LAWRENCE, D. H. The nature of a stimulus. In S. Koch (Ed.), *Psychology: a study of a science.* Study II. Vol. 5. *Process Areas.* New York: McGraw-Hill, 1963.

MCCULLOCH, T. L., and PRATT, J. G. A study of the pre-solution period in weight discrimination by white rats. *Psychol. Rev.*, 1934, **18**, 271–290.

OSGOOD, C. E. *Method and theory in experimental psychology.* New York: Oxford Univer. Press, 1953.

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

RESTLE, F. A theory of discrimination learning. *Psychol. Rev.*, 1955, **62**, 11–19.

RESTLE, F. Discrimination of cues in mazes: a resolution of the "place-vs-response" question. *Psychol. Rev.*, 1957, **64**, 217–228.

RESTLE, F. Statistical methods for theory of cue learning. *Psychometrika*, 1961, **26**, 291–306.

RESTLE, F. The selection of strategies in cue learning. *Psychol. Rev.*, 1962, **69**, 329–343.

SCHARLOCK, D. P. The role of extramaze cues in place and response learning. *J. exp. Psychol.*, 1955, **50**, 249–254.

SHEPARD, R. N., HOVLAND, C. I., and JENKINS, H. M. Learning and memorization of classifications. *Psychol. Monogr.*, 1961, **75**, No. 13 (Whole No. 517).

SIEGEL, S. *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill, 1956.

SPENCE, K. W. Continuous versus noncontinuous interpretations of discrimination learning. *Psychol. Rev.*, 1940, **54**, 223–229.

SUPPES, P., and GINSBERG, ROSE. A fundamental property of all-or-none models; binomial distribution of responses prior to conditioning, with application to concept formation in children. Psychology Series, Technical Report No. 39, Institute for Mathematical Studies in the Social Sciences, Stanford Univer., 1961. (Published in *Psychol. Rev.*, 1963, **70**, 139–161.)

TRABASSO, T. R. The effect of stimulus emphasis on the learning and transfer of concepts. Unpublished doctoral dissertation, Michigan State Univer., 1961.

TRABASSO, T. R. Stimulus emphasis and all-or-none learning in concept identification. *J. exp. Psychol.*, in press.

UNDERWOOD, B. J., and RICHARDSON, J. Some verbal materials for the study of concept formation. *Psychol. Bull.*, 1956, **53**, 84–95.

UNDERWOOD, B. J., and SHULZ, R. W. *Meaningfulness and verbal learning*. Philadelphia: Lippincott, 1960.

WARREN, J. M. Stimulus perseveration in discrimination learning by cats. *J. comp. physiol. Psychol.*, 1959, **52**, 99–101.

WYCKOFF, L. B. The role of observing responses in discrimination learning. Pt. I. *Psychol. Rev.*, 1952, **59**, 431–442.

ZEAMAN, D., HOUSE, B., and FONDA, C. *An attention theory of retardate discrimination learning*. Progress Report No. 3, NIMH USPHS 1961 Res. Grant M-1099, Univer. of Connecticut.