J. Hintikka & P. Suppes (Eds.), Aspects of Inductive Logic. Amsterdam: North-Holland, 1966. pp 21-48.

CONCEPT FORMATION AND BAYESIAN DECISIONS*

PATRICK SUPPES
Stanford University, Stanford, California

1. Introduction. The primary aim of this paper is to examine and develop some relations between decision theory and recent work on concept formation by learning theorists. Some of the ground rules of this investigation perhaps need to be stated at the very beginning. Let me first try to make clear how I conceived in a general way the relation between decisions and concept formation. If we examine the structure of decision theory as expounded for example, in the excellent book of Savage [1954], we find that there is really no place for the formation of new concepts by the decision-maker. The theory is conceived in such a way that the decision-maker has a probability distribution over all possible states of the world and a utility function over all possible future histories of the universe. As observations are made or experiments performed by the decision-maker, the information received is brought into his formal decision framework by appropriate modifications of initial probabilities as new conditional probabilities. The important thing I wish to emphasize is that the theory provides no place for the decision-maker to acquire a new concept on the basis of new information received. The theory is static in the sense that it is assumed the decision-maker has a fixed conceptual apparatus available to him throughout time.

There are, it seems to me, two important ways in which concept formation enters in the making of actual decisions. The first kind of modification in the decision structure that may be introduced by concept formation is a relatively straightforward refinement or at least modification of the initial partition of the possible states of nature by the consideration of additional concepts. The consideration of these additional concepts is almost always brought about by the reception of a cue or stimulus resulting from some new observation. The essential thing however in this kind of modification is that the concepts newly introduced are already a part of the conceptual apparatus of the decision-

* The work on this paper was supported by a contract between ARPA, U.S. Department of Defense and the System Development Corporation.

maker. It is just that he has not been using them to partition the space of the states of nature until a particularly critical, or new, sort of observation was obtained. Because the new concept brought into focus is actually one already known to the decision-maker, it is becoming customary in the psychological literature to call this process concept identification, rather than concept formation, and we shall so in fact refer to it here.

Advocates like de Finetti and Savage of Bayesian theory would indeed claim that this first kind of concept formation is already taken care of by considering all the possible states of the world and all possible future histories. From a theoretical standpoint, or at least one theoretical standpoint, there is indeed a good argument to back up this claim. Yet from a more behavioristic viewpoint it is quite unrealistic, for no actual decision-maker is able in any genuine way to define an a priori distribution over all possible states of the world or a utility function over all future histories. His powers of discrimination and analysis, even in terms of the empirical data available to him, are inadequate to this task. In actual practice the decision-maker is always operating with what Savage has termed a small-world situation. The decision-maker operates with a fairly small number of concepts and the partition of the possible states of nature generated by these concepts. I want to emphasize that it is not necessary that the partition itself be finite for some of the concepts may be conceived by the experimenter as being measured on a continuum. The crucial thing is that the concept space is always finite-dimensional and in fact, the number of dimensions is a relatively small integer.

The second way in which concept formation modifies the decision structure is the genuine case of concept formation proper. In this instance the decision-maker actually forms a concept he did not previously have in his repertoire. Numerous examples of this kind of concept formation are to be found in the learning experience of anyone. It may be rightly claimed that this kind of concept formation is essential to any major advance in science or technology.

In the next section I turn to some simple examples of concept identification and attempt to show how they disturb the simple Bayesian picture of decision making. In the following section I consider some common problems besetting Bayesian and stimulus-sampling learning models. In the final section I sketch a possible line of attack on the structural or combinatorial problems facing any theory of concept formation. I also try to show that Bayesian considerations are not central to the most pressing problems of the theory of concept formation, and that no theory of complex problem solving is possible without an approximate solution to these problems.

Before embarking on these somewhat detailed considerations, I would like to indicate in a general way how the subject matter of this paper relates to the more standard literature of inductive logic. The most important and also the most subtle point centers around the conception of rational behavior back of the general criteria used to evaluate an inductive logic or procedure. In the case of deductive logic the response is simple and clear. The criteria of soundness and completeness make no allowance for an imperfect or limited knower. The inattention to the obvious finite capacity of any actual knower is a simplifying abstraction that makes the mathematical theory of deductive inference a manageable subject in the tradition of classical mathematics.

To a large extent the same sort of simplifying abstraction has been assumed in deductive logic, but with an important difference. No adequate inductive criteria corresponding to the deductive criteria of soundness and completeness are as yet available. Bayesian decision theory provides a possible answer, but certainly not one that is as yet uniformly acceptable. In my own judgment the problem of finding such criteria in inductive logic is not as interesting as in deductive logic, because the finite capacity of the learner (when talking about induction it seems more natural to speak of a *learner* rather than a *knower*) is central to the fundamental problem of making an induction from a finite sample. Put another way, problems of induction seem continually to run up against massive combinatorial problems that do not play the same essential role in deduction. And once we begin talk about, say, $(2)^{2^{10}}$ possibilities it is natural to ask about the kind of learner that is going to "look over" possibilities whose number is of this order of magnitude, and then to try to inject some semi-realism into the discussion. Now that it is generally recognized that even the biggest conceivable computers could not attack by brute force methods the combinatorial problems of playing a winning game of chess, for example, the crucial role of concept formation in providing a powerful method of introducing new and necessary structure is more easily made apparent.

In addition, the continued concern in the literature of inductive logic with overly simplified, unrealistic problems suggests that there is a useful place for an explicit analysis of why even the relatively powerful Bayesian methods of induction are far too weak to solve most complex problems.

From a more general standpoint, then, an objective of this paper is to make a contribution to the analysis of the concept of rationality. The discussions of rationality in the literature of induction or ethics seem to have largely ignored the difficult problems of concept formation that must be faced by any agent that does not have an unlimited memory and unlimited powers of analysis.

2. Some simple examples of concept identification. To illustrate some of the comparisons I want to draw between a Bayesian approach to information processing and decision making, on the one hand, and psychological models of behavior on the other, I shall begin with an experiment that is really too simple to be described as a concept experiment, but because of its very simplicity will be a satisfactory paradigm for the making of certain initial distinctions.

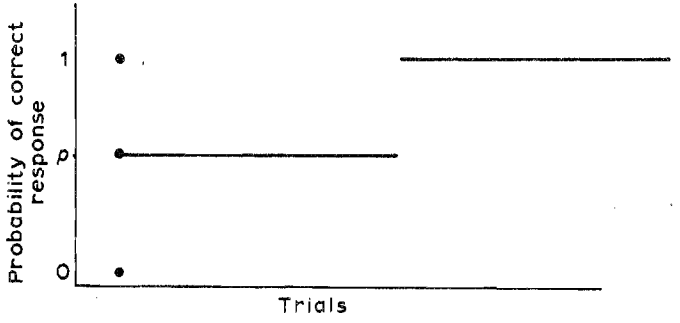
I have in mind a simple paired-associate experiment. The task for the subject is to learn to associate each one of a list of nonsense syllables with an appropriate response. In a typical setup the list might consist of twenty nonsense syllables of the form CVC. The responses are given by pressing one of two keys. On a random basis ten of the syllables are assigned to key 1 and ten to key 2. The subject is shown each nonsense syllable in turn, is asked to make a response, and is then shown the correct response by one of several devices, for example, by the illumination of a small light above the correct key. After the subject has proceeded through the list once, he is taken through the list a second time but the order of presentation of the twenty items is randomized. A criterion of learning is set, for example, four times through the list without a mistake. The subject is asked to continue to respond until he satisfies this criterion. The criterion is selected so as to give substantial evidence that the subject has indeed learned the correct association of each stimulus item and its appropriate response – at least this language of association is the one ordinarily used by many psychologists concerned with this type of experiment.

Let me describe two simple psychological models for this experiment before discussing the obvious Bayesian model and its defects. The simple stimulus-association model to be applied to the phenomena is the following. The subject begins the experiment by not knowing the arbitrary association established by the experimenter between individual stimuli and the response keys. He is thus in the unconditioned state U. On each trial there is a constant probability c that he will pass from the unconditioned state to the conditioned state C. It is postulated that this probability c is constant over trials and independent of responses on preceding trials. Once the subject has passed into the conditioned state it is also postulated that he remains there for the balance of the experiment. A simple transition matrix for the model, which is a first-order Markov chain with two states U and C, is the following:

	C	U
C	1	0
U	c	$1-c$

To complete the model for the analysis of experimental data it is also necessary to state what the probabilities of response are in the two states U and C. When the subject is in the unconditioned state, it is postulated that there is a guessing probability p of making a correct response, and that this guessing probability is independent of the trial number and the preceding pattern of responses. When the subject is in the conditioned state, the probability of making a correct response is postulated to be 1.

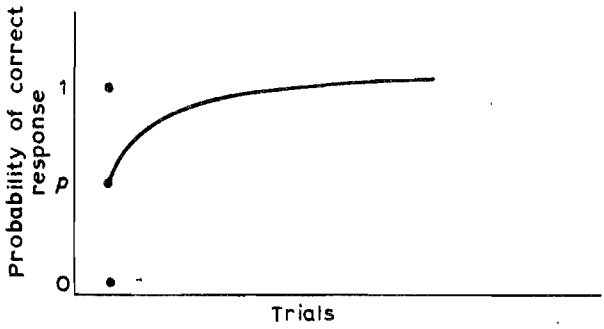
The most striking psychological aspect of the stimulus-association model just described is the all-or-none character it postulates for the learning process. The organism responds with a constant guessing probability until the correct conditioning association is established on an all-or-none basis. From that point on he responds correctly with probability 1. This means that for an individual subject the individual learning curve has the following simple appearance.



The important thing to note about this curve is that it is perfectly flat until conditioning occurs and at that point there is a strong discontinuity.

The second psychological model is a linear-incremental model that postulates that the probability of making a correct response increases each time the subject is exposed to the stimulus and is shown the correct response. Let p_n be the probability of a correct response on trial n, and let $q_n = 1 - p_n$, that is, let q_n be the probability of an incorrect response or error on trial n. The simplest way of formulating this model is in terms of q_n. It is postulated that the following recursion will describe the course of learning: $q_{n+1} = \alpha q_n$. This linear model can be put within the framework of stimulus-association theory in a rather simple way. Instead of postulating that a single stimulus is being sampled and conditioned in connection with each nonsense syllable displayed, it may be postulated that there are a large number of stimuli being sampled and conditioned. These are simple and reasonable assumptions

about sampling and conditioning. As the number of stimuli becomes quite large, the linear model emerges as an asymptotic limit. (For a detailed derivation of the linear model from stimulus-sampling and conditioning assumptions, see Estes and Suppes [1959].) The learning curve postulated for an individual subject by the linear models looks something like the following.



As is evident enough from the two theoretical learning curves for individual subjects predicted by the two models, there are quite sharp behavioral differences in the predictions of the one-element stimulus-association model and the linear-incremental model. On the other hand, it is worth noting that the matter of discriminating the two models must be approached with some care. For example, the mean learning curve obtained by averaging data over a group of subjects, or a group of subjects and a list of items as well, is precisely the same for the two models. In the linear model it would naturally be written:

$$q_n = \alpha^{n-1} q_1 . \tag{1}$$

In the one-element stimulus-association model the same mean learning curve would naturally be written:

$$q_{n+1} = (1 - c)^{n-1} q . \tag{2}$$

In estimating parameters from behavioral data it is natural to equate p_1 and p (or q_1 and q) and, on that basis, the estimate of α will simply be the same as the estimate of $1-c$; there is no behavioral difference between the two models in the prediction of the mean learning curve. On the other hand, perhaps the most striking difference between the two models can be obtained by looking at data prior to the last error, that is, we sum data over subjects and items, but we restrict that summation to response data occurring before

the last error on a given subject-item. When data are summed in this fashion, the one-element stimulus-association model predicts the discontinuous learning curve shown above for an individual subject, whereas the linear-incremental model predicts a smooth incremental learning curve. That the data from experiments of this kind favor very much the one-element stimulus-association models over the linear-incremental models has been shown by a number of experiments (see, e.g., Bower [1961]).

Let us now attempt to apply Bayes' theorem in a correspondingly direct way to an analysis of the paired-associate experiment. Without any loss of generality we may restrict the analysis to a single item, that is, to the learning of a single association between a given nonsense syllable and the correct response. Let H_1 be the hypothesis that the correct response for the single syllable is response 1, and let H_2 be defined similarly. It is natural to assume that the a priori probabilities $P(H_1)$ and $P(H_2)$ are each a half. (As we shall see, for this simple situation the particular assumption made about the a priori probabilities is of no real importance.) In the present case, the "evidence events" are easy to describe and amount essentially to a complete confirmation of one of the two hypotheses. Let us define the evidence event E_i as the event of being shown that the nonsense syllable is associated with the response i. It should be clear how to define the conditional probability $P(E_j|H_i)$ which is called the *likelihood* of H_i when E_j is observed. In the present simple case the likelihoods must either be 0 or 1. The likelihoods are 1 when $i=j$, and 0 when $i\neq j$. We may then compute the a posteriori probabilities $P(H_i|E_j)$ according to the usual Bayes formula:

$$P(H_i \mid E_j) = \frac{P(E_j \mid H_i)\,P(H_i)}{\sum_i P(E_j \mid H_i)\,P(H_i)}.$$

Again in the present case the computation of these a posteriori probabilities is simple and immediate. If $i=j$ the a posteriori probability is 1 and if $i\neq j$ the a posteriori probability is 0.

Thus,

$$P(H_1 \mid E_1) = \frac{1\cdot\frac{1}{2}}{1\cdot\frac{1}{2}+0\cdot\frac{1}{2}} = 1$$

$$P(H_1 \mid E_2) = \frac{0\cdot\frac{1}{2}}{0\cdot\frac{1}{2}+1\cdot\frac{1}{2}} = 0$$

$$P(H_2 \mid E_1) = 0$$

$$P(H_2 \mid E_2) = 1.$$

Note that the results are independent of the a priori probabilities $P(H_1)$ and $P(H_2)$ as long as these probabilities are in the open interval $(0,1)$.

In the present case, then, how is the application of a Bayesian approach related to the two psychological models already sketched for the learning process? The answer I think is obvious. The one-element stimulus-association model yields exactly the same predictions as the Bayes model, if it is assumed that the conditioning parameter c has the value 1, that is, if it is assumed that the subject always learns in one trial a correct association between nonsense syllable and response key. There are too many experiments to need detailed citations here to show that the assumption that $c = 1$ is not a reasonable one for paired-associate experiments. There is no doubt that the Bayesian model does not provide a good account of actual behavior in these experiments. Its generalization in the form of the one-element model is much more satisfactory.

Advocates of a Bayesian approach as the first approximation to actual behavior will be quick to retort that they have in mind the application of the model to situations in which the subject can utilize his full resources of memory and reasoning. Some may wish to point out that indeed c would equal 1 if the subject were permitted to use pencil and paper in the course of the experiment, and simply to write down the correct association between the stimulus and response once it has been shown to him. However, any serious consideration of the general purpose and intention of such a paired-associate experiment quickly shows that this defense of the Bayesian approach as an explanatory model of actual behavior is not really satisfactory at all. The paired-associate experiment is defined and set up in the manner that it is in order to provide an extremely simple paradigm of learning. The simplicity of that paradigm is destroyed once a subject is permitted such recording devices as a pencil and paper. A scientific hope of such experiments is that an adequate fundamental theory of the learning process can be developed for learning stripped of complicated processes of memory, association and reasoning that are utilized in everyday decisions. If the fundamental theory is genuinely correct, then it will lead ultimately to extensions to more complicated situations including the sort in which the learning problem confronting the subject is not one that he can trivialize by the use of some additional simple devices. Some of my subsequent examples will in fact be instances of this kind of experiment.

A second kind of objection that might be offered by Bayesians to a comparison of the three models is that the real purpose of the Bayesian approach is to prescribe a normative course of behavior and not describe actual

behavior. In spite of the persuasiveness of the important distinction between normative and descriptive theories, this argument is too facile by half. The kind of situations which decision-makers are continually confronted with is precisely the kind of situation in which we place the subject. The subject faced with the paired-associate problem could, if he were given paper and pencil, readily and quickly solve the problem, but the point is that he is not given these additional aids. The decision-maker, whether it be an executive faced with a major policy decision, a logistics expert deciding on the next quarter's inventory, or a legislator deciding on how to present a crucial and controversial bill, is in a situation analogous to that of our subject, for the complexities of the one correspond to the simple restrictions of the other. In certain cases, given unlimited budget for computing purposes and unlimited staff to furnish scientific information, it might be possible for the decision-maker to act rather completely like a Bayesian strategist. When this is not possible, as it usually is not, the decision-maker must make a large number of rough and ready judgments that do not easily fit within the frame of a detailed normative theory. Indeed, a primary aim of this paper is to show by the consideration of several simple examples that the attempt to structure the decision-making process entirely within the Bayesian framework will lead to serious miscalculations about actual performance and, in certain cases, to bad advice on normative performance.

I now turn to a first simple example of concept identification. Let us suppose that a subject is to be shown triangles of various sizes, and let us also suppose that the instructions are meant to bias him in the direction of paying attention to size only. We tell him that he is to classify the triangles primarily according to size into Class A or Class B. In actual fact, in addition to having triangles of three different areas, each triangle will have the property of having an angle less than 15° or no angles less than 22½°. Let us call the three sizes a, b and c and the two angle properties s and t. Suppose we fix that Class A will consist of the triangles with the property a–s and b–t. Class B will then consist of the complement of Class A, that is, of the combinations a–t, b–s, c–s and c–t. I have picked the angle property because it is a feature of triangles that does not have much saliency for untrained subjects, whereas size is ordinarily a highly salient property. With these instructions and in this situation it is quite probable that many subjects would have a Bayesian distribution of prior probabilities that are non-zero only in terms of hypotheses about size. It does not really matter what specific prior probabilities we assume on hypotheses about size. Eliminating the hypothesis that all sizes belong in Class A and also the hypothesis of no sizes in Class A, there are six

size hypotheses remaining, which we may write in terms of sizes assigned to Class A, as H_a, H_b, H_{ab}, H_c, H_{ac} and H_{bc}. We may assume a positive probability for each of the six. It is also a condition of the experiment that the six possible types of triangles are presented by the experimenter on an equally likely basis. It is easy to see that the hypotheses H_a, H_b and H_{ab} will each have a probability of two-thirds of being correct. The explicit situation is shown in the following table, where the entry 1 indicates a correct classification, and 0 an incorrect classification, under each of the six hypotheses for the six types of figures.

Type of figure		Hypotheses					
		H_a	H_b	H_{ab}	H_c	H_{ac}	H_{bc}
A	*a–s*	1	0	1	0	1	0
B	*a–t*	0	1	0	1	0	1
B	*b–s*	1	0	0	1	1	0
A	*b–t*	0	1	1	0	0	1
B	*c–s*	1	1	1	0	0	0
B	*c–t*	1	1	1	0	0	0

In the table the four cases for which each of the three hypotheses H_a, H_b and H_{ab} is correct are indicated. The remaining three hypotheses, namely, H_c, H_{ac} and H_{bc}, will asymptotically each have a probability 0. Within the framework of the six hypotheses the subject can do no better than indifferently select among H_a, H_{ab} and H_b.

It is clear from this analysis that from the Bayesian standpoint any subject who begins with his entire prior distribution weighted on the six hypotheses concerned with size alone will not be able to solve the problem completely. It may of course be objected that the assumption made about prior probabilities is not a reasonable one. The issue is complicated and I do not mean to suggest that I think definitive arguments can be given in support of the kind of assumption made. There is, however, a certain amount of evidence both from the behavior of subjects, and interrogation of them about their behavior, to show that in experiments in which an ultimately relevant concept has a very small degree of saliency, the subject begins initially by completely ignoring this concept or property. For such situations there would seem to be

only a Pickwickean sense in which a strictly positive distribution over hypotheses involving this concept can be postulated.

Let us now consider how we would approach the analysis of the subject mastering the problem in terms of some of the ideas of concept formation that have been developed in the last couple of years. The theoretical account I shall give will be somewhat more elaborate than the models much tested in recent experiments on concept formation (as for example, Bourne and Restle [1959], Bower and Trabasso [1964] and Suppes and Ginsberg [1962a], [1962b], [1963]).

As the first stage of learning, let us assume that the subject, following the verbal cue given him by the experimenter, samples only the three size stimuli, which we have designated as a, b and c. Initially he does not know how each of these stimuli should be connected or associated with Class A or Class B. We may thus postulate that they are in the unconditioned state. When a stimulus presentation is given which permits a sampling of one of the three stimuli, then in terms of the correction procedure given, that is, the statement as to whether or not the figure shown belongs to Class A or Class B, we may postulate a probability c that the size stimulus sampled will become conditioned to one of the two classes, that is, to one of the responses A or B. Notice that we are postulating initially that the subject samples with probability 1 whichever one of the size stimuli is available on a given trial. On this basis, the learning for stimulus c is particularly simple. We may just apply the one-element model described above for paired-associate learning. This stimulus starts in the unconditioned state and with probability c on each occasion on which it is sampled it enters the conditioned state, in this case, conditioning to Class B or Response B. When stimulus c is conditioned to Response B, on every occasion in which Response B is made on the presentation of this stimulus, the classification is proved to be correct, and therefore there are no grounds for the subject's changing or modifying this conditioning. In a complete sense, the conditioning of stimulus c exemplifies, as postulated here, the one-element model for paired-associate learning described above. (The use of 'c' to refer both to the stimulus and its probability of conditioning should not be a source of confusion, as the specific reference intended is always clear from the context.)

The situation is considerably more complicated for stimuli a and b. Half the time that stimulus a or b is sampled the presented figure that has size property a or b can be classified in Class A, and the other half of the time, on a random basis, Class B. Intuitively when the subject finds that he cannot use stimulus a or b to make a correct classification, he will be led to sample other

properties or aspects of the figure. Before he is led to make this additional sampling, there is often one strategy he will try. He may judge that his initial response connection for one of the stimuli was incorrect and he will reverse the association. For example, if on the first occasion that stimulus b is sampled, it turns out that the figure is classified as Class A, but on the second occasion that stimulus b is sampled, the figure is put in Class B by the experimenter, he may reverse the association and not yet be led to sample other stimuli. Let us designate the probability of such a reversal of the association by r, and let us postulate that he will sample a new property with probability s, when it turns out that the association that he has established is wrong. Extending the kind of assumptions that went into the development of the one-element model for paired-associate learning, we may postulate a four-state Markov process describing this stage of the subject's learning. He begins in state U representing the fact that stimulus a, let us say, is unconditioned. We may pass from state U to either state A or B representing two possible responses to which stimulus a may be conditioned. After reaching state A or B, he will on each trial that a is sampled be incorrect with probability $\frac{1}{2}$. The matrix is then constructed so as to postulate that with probability $\frac{1}{2}s$ he enters state N, the state in which he samples a new property, and with probability $\frac{1}{2}r$ he reverses the stimulus association from response A or response B or vice versa as the case may be. It is of course a constraint of the model that $r+s \leqq 1$. The complete matrix then is as follows:

	N	B	A	U
N	1	0	0	0
B	$\frac{1}{2}s$	$1-\frac{1}{2}s-\frac{1}{2}r$	$\frac{1}{2}r$	0
A	$\frac{1}{2}s$	$\frac{1}{2}r$	$1-\frac{1}{2}s-\frac{1}{2}r$	0
U	0	$\frac{1}{2}c$	$\frac{1}{2}c$	$1-c$

Note that state N is the absorbing state of this chain, because we are postulating that the subject will always be led, on the basis of his failure with the association established for the size stimuli a and b, to the sampling of a new property. It should also be remarked that this matrix represents the situation for stimulus b as well as for stimulus a. We are, as in the fashion of paired-associate learning, postulating that the process of being led to state N when stimulus a is sampled is statistically independent of the process of being led to state N when stimulus b is sampled. No doubt in actual practice this assumption is probably slightly violated but it makes the quantitative treatment of the concept identification considerably more simple, and is therefore,

a desirable feature of a first approximation. For fast learners we may postulate that $s=1$ and $c=1$. The matrix then assumes the following simple form:

	N	B	A	U
N	1	0	0	0
B	$\frac{1}{2}$	$\frac{1}{2}$	0	0
A	$\frac{1}{2}$	0	$\frac{1}{2}$	0
U	0	$\frac{1}{2}$	$\frac{1}{2}$	0

In a given experiment more detailed knowledge may be obtained by looking at the actual sequence of presentations of figures and observing their classification. If, for example, the subject was always wrong in classifying a figure with stimulus a in initial trials, he could be in state N on the third trial that a figure with size property a is presented. It may also be noted that we have postulated that in this concept-identification task the subject is learning only on trials on which he makes an error. Bower and Trabasso [1964] present impressive evidence that for the kind of experiment described here this is roughly the situation. I shall have more to say on this point later.

Upon entering state N the subject is now in a position to sample a new property. Note the difference from the Bayesian formulation. Up to this point the probability of sampling any property other than a size stimulus has been 0. It is only due to the failure of the size stimuli to lead to the correct solution that the subject has been forced to change his initial distribution and sample other properties. Suppose for instance that the subject now samples stimuli connected with the orientation of the base of the triangle. We may suppose that the base of the triangle varies from the horizontal in three different angles, namely 0°, 15° and 30° (these numerical values are taken for purposes of illustration only). We also shall suppose that the occurrence of figures with these respective orientations is randomly assigned independent of other characteristics and therefore any particular orientation will occur in Class A figures approximately half of the time and in Class B figures the other half. Various things can be postulated at this point. We can assume that the subject disregards size stimuli entirely and samples only orientation stimuli, or we can also postulate that he samples a size-orientation pattern combining stimuli exhibiting both properties. For many situations this latter pattern assumption has been shown to be a sound one. But whichever sampling procedure he adopts at this point, that is, concentration only on orientation or pattern sampling of orientation and size together, he will be led to the same results as before and will once again enter state N, and be required to select a new property for sampling.

Parenthetically it may be remarked that for those readers who fear that the process of concept identification as described here is too slow to describe what actually takes place, it may be said that it is not difficult to cite experiments in which a large number of trials is required by subjects to master what may appear to an experimenter or an observer with full knowledge of the situation as absurdly simple identification problems. When the number of trials to complete mastery of the problem is on the order of a hundred, many opportunities are presented for sampling different properties of the stimulus display presented.

There are two important factors I have ignored in this analysis but which would in all likelihood enhance the rate of learning or the rate of concept identification. One is the factor of memory. When a new property is sampled, in many cases it is sampled and rejected simply on the basis of its ability to account for correct classification of items already seen and whose classification is remembered. On the other hand, it is not an unrealistic assumption to suppose that the transition matrix described above is essentially the sort of one that is used in testing from memory new sampled properties. The experimental difficulty of course is that it is not a simple or direct matter to elicit behavioral data giving evidence on this point.

The second related phenomenon that I have ignored is the undoubted fact that when a given property is being sampled and used as a basis of classification it is often the case that simultaneously other properties are being sampled and silently rehearsed, meaning by this that their ability correctly to classify is being noticed even though they are not the properties used by the subject in making his classification on the given trial. Again it is not unreasonable to suppose that the process of rehearsal may be represented by a transition matrix very similar to the one given above. There is considerable indirect experimental evidence of the efficacy of rehearsal from the standpoint of learning. Several experimental studies have shown the positive effects of an increased amount of study time on the rate of learning of paired-associates.

As I understand the matter, no simple Bayesian approach to information processing and decision making would take explicit account of these two aspects of concept formation and learning, namely, the effects of memory and rehearsal. It should be emphasized that in another sense memory may be taken account of in Bayesian procedures. Modern empirical Bayes procedures have in many cases been developed on the assumption of a finite memory, but the kind of use of memory suggested here is of a different sort, namely, memory of what happened on preceding trials is used in a new way on trial n to check out the efficacy of a property not considered or sampled prior to

trial n. From the Bayesian standpoint the use of this property on trial n in terms of items from memory would require the assumption that the property or concept had a positive prior distribution on earlier trials.

The relatively simple concept-identification problem we have been considering is already beyond the resources of the standard systems of inductive logic, because the subject in the experiment is not told what are the relevant elementary properties. Although in principle inductive logics of the Carnapian variety have a method for handling questions of relevance, in practice they do not deal with the kind of thing that arises in any concept-identification experiment when the subject is not told what is relevant – to pick a very simple example, the relevant aspects might turn out to be relations rather than properties. Also, methods of constructing such logics as matters now stand do not provide any guidelines for enumerating large sets of properties among which the relevant ones are likely to lie.

Focusing on concept-identification experiments makes it possible to draw an important distinction between Bayesian theory and the Carnapian sort of inductive logic. To make the standard inductive logic apply it is necessary to codify explicitly in the language of the logic all the evidence of past experience the subject considers pertinent to the experiment, but this I would claim is always a hopeless task. It is difficult enough to narrow the situation down to a manageable set of properties and relations, but it is humanly impossible to lay out all the evidence that went into the selection of this set and the beliefs held about its members. To put it in simplest terms, it is at the least the problem of having a limited, finite memory. The Bayesian approach, on the other hand, is not bedeviled by this difficulty, because past experience can be encoded in the a priori distribution over the selected set of properties and relations. Once again, it is a question of a realistic conception of rationality. If we want to explicate the concept of rational human behavior, and not that of omniscient rational behavior, limitations on memory and computing power must be taken seriously. Taking such limitations seriously is of course imperative in attempting to apply an inductive logic. (The fact that these limitations are fundamental is why within the domain of deductive logic the theory of recursive functions is of quite restricted use in theorizing about or applying actual computers.)

3. Some common problems of Bayesian and stimulus-sampling models. The discussion of the last section to a certain extent overemphasizes the differences between Bayesian and stimulus-sampling models for decisions, particularly when the decisions involve concept identification. In the present section I

want to emphasize some of the commonality between the two kinds of models and to point out some of the problems that beset them both. A particular point of this section is to show that many of the differences often emphasized in discussions of the cognitive or Bayesian approach as opposed to the stimulus-response approach is a difference primarily in terminology, and not so much in something that is sharply defined and empirically observable.

A convenient place to begin is with the classical case of a two-choice problem with noncontingent reinforcement. The problem for the subject on each trial is to predict which one of two lights will flash. Using familiar notation, let us call E_1 the reinforcing event corresponding to the flashing of the left light and E_2 the reinforcing event corresponding to the flashing of the right light. The response that consists of predicting that the left light will flash is designated A_1 and the response that consists of predicting that the right light will flash is designated A_2.

The noncontingency of the situation is defined by making the probability of an E_1 reinforcement on each trial equal to π and the probability of an E_2 reinforcement $1-\pi$. It is understood that the events E_1 and E_2 are mutually exclusive and exhaustive, that is, on each trial exactly one of the two lights flashes, and the probability of which will flash is fixed by the parameter π. Everything that we have to say in what immediately follows applies, mutatis mutandis, to other more complicated reinforcement schedules, but the basic principles are precisely the same.

Let us begin by considering some Bayesian models for this situation. In the first place these Bayesian models shall be defined in terms of several sets of hypotheses, and we shall call an exhaustive set of hypotheses, that is a set of hypotheses that covers every contingency, a *strategy*. It is understood that in the ordinary Bayesian terminology what I am now calling strategies would very often be called hypotheses, but the present language is suggestive of game-theoretic language, as well as of the kind of language that has been used by various people interested in cognitive models of the learning process. For simplicity let us begin with four hypotheses:

h_1 : an E_1 reinforcement is followed by an E_1;
$h_{1'}$: an E_1 reinforcement is followed by an E_2;
h_2 : an E_2 reinforcement is followed by an E_2;
$h_{2'}$: an E_2 reinforcement is followed by an E_1.

Given the above four hypotheses and the fact that h_1 and $h_{1'}$ (and h_2 and $h_{2'}$) are incompatible, a strategy for the subject consists of believing, or acting as if he believed, one of the following four pairs of hypotheses: (h_1, h_2),

$(h_1, h_{2'})$, $(h_{1'}, h_2)$, $(h_{1'}, h_{2'})$. Thus the strategy (h_1, h_2) requires that an A_1 response will be made if on the preceding trial an E_1 reinforcement occurred, and an A_2 response will be made if an E_2 reinforcement occurred on the preceding trial.

As is apparent from what has already been said, the four strategies correspond to the four hypotheses relevant in the sense of a Bayesian model. Granted only a positive a priori probability for each of the four strategies, it is clear what is the asymptotic prediction of the Bayesian model, that is, what the asymptotic a posteriori probabilities of the strategies will be. Namely,

$$\begin{aligned} P(h_1, h_2) &= \pi(1-\pi) \\ P(h_1, h_{2'}) &= \pi^2 \\ P(h_{1'}, h_2) &= (1-\pi)^2 \\ P(h_{1'}, h_{2'}) &= (1-\pi)\pi . \end{aligned}$$

The Bayesian decision-maker with unlimited memory will then choose strategy $(h_1, h_{2'})$ with probability 1 for $\pi > \frac{1}{2}$.

On the face of it this familiar Bayesian result leading to selection of event E_1 with probability 1 seems very much in conflict with the standard theoretical results obtained in stimulus-sampling theory, which predicts that an A_1 response will be made as a prediction of an E_1 reinforcement with asymptotic probability π, and we have the well-known matching law, first formulated by W. K. Estes.

From what has been said thus far it is easy enough to formulate the stimulus-sampling model with N stimulus elements in the noncontingent situation. On each trial the organism is sampling one stimulus. It becomes conditioned to the response that is reinforced with probability c, and with probability $1-c$ its conditioning does not change. Among the N stimuli exactly one is sampled on each trial. This sampling takes place on a random basis, that is, there is a probability $1/N$ of any particular stimulus' being sampled, independent of what else may have occurred on past trials. When a stimulus is sampled the response is made to which that stimulus is conditioned. In terms of these theoretical assumptions, the behavior of the subject may be defined in terms of the parameters c, N and π.

This description of stimulus-sampling theory seems quite different from the Bayesian approach. I now want to show how closely related they actually are, and how easily a formal isomorphism between models of the two theories may be set up. To bring the two together let us first examine the Bayesian model under a highly restricted memory assumption. In particular,

let us suppose that the subject, although he is a Bayesian, is only able to remember what happened the last time a test of any particular hypothesis in his strategy was made. Whenever the outcome of this test is negative, he immediately changes his strategy by replacing the incorrect hypothesis by the correct one. For example, if his strategy is (h_1, h_2) and he finds on trial n that the E_1 reinforcement occurring on trial $n-1$ is followed by E_2, he then immediately changes his strategy to $(h_{1'}, h_2)$. In other words, he is making no use of any evidence concerning trial pairs for which an E_1 reinforcement is followed by an E_1 before trial n. His memory is of minimal length to use any of the evidence relevant to the hypotheses at all.

The transition matrix for this Bayesian model with memory of length one is then the following:

	$(h_1, h_{2'})$	(h_1, h_2)	$(h_{1'}, h_{2'})$	$(h_{1'}, h_2)$
$(h_1, h_{2'})$	π	$(1-\pi)^2$	$\pi(1-\pi)$	0
(h_1, h_2)	$\pi(1-\pi)$	$1-2\pi(1-\pi)$	0	$\pi(1-\pi)$
$(h_{1'}, h_{2'})$	π^2	0	$1-(\pi^2+(1-\pi)^2)$	$(1-\pi)^2$
$(h_{1'}, h_2)$	0	π^2	$\pi(1-\pi)$	$1-\pi$

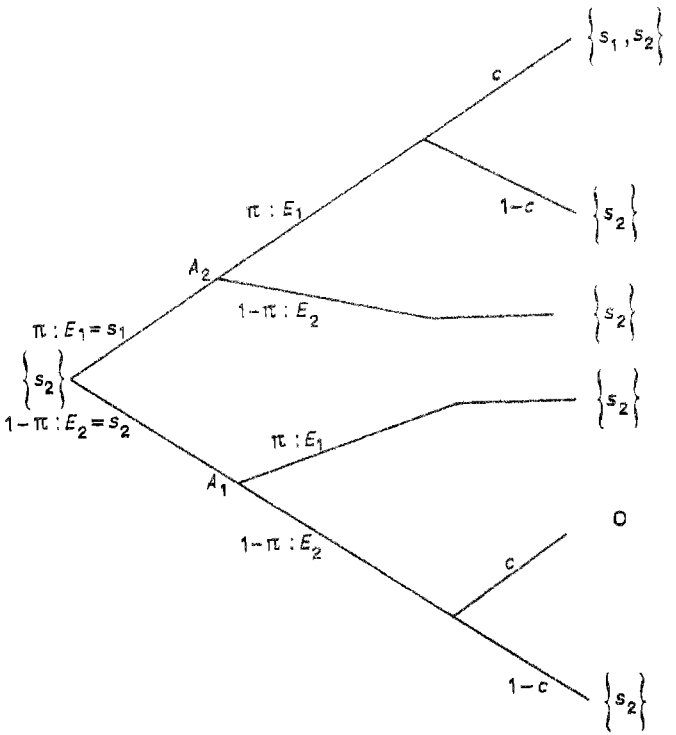
It is a simple matter to show that the asymptotic probability of an A_1 response with this transition matrix and the states as defined above is that of the familiar matching law, namely, π.

Let us now look at a formulation of the analog of the Bayesian model in terms of stimulus-sampling theory. As has already been indicated, its theoretical assumptions are formulated along the following lines. There is available a set of N stimuli which the subject has conditioned or associated with various possible responses. At the beginning of a trial the organism is in a certain state of conditioning. A set, possibly a proper subset of the N stimuli, is presented to him and he samples on a random basis, that is, with a uniform distribution. Exactly one of the stimuli is sampled and then the organism responds in terms of the association bond that stimulus has with one of the possible responses (in case the stimulus sampled is not conditioned to any response a guessing response is made of the kind already described in discussing paired-associate learning). After the response is made, the reinforcement is given and the stimulus sampled changes its conditioning with probability c to the response reinforced, in case the response made was incorrect.

To apply these postulates and develop a model corresponding to the Bayesian model we shall assume there are exactly two stimuli. One of them is the E_1 reinforcing event occurring on the preceding trial and the other is the

E_2 event occurring on the preceding trial. Thus, on each trial the subject has available exactly one of the two stimuli to sample, and the whole sampling process is thereby trivialized. Quite apart from this identification of the two stimuli, on the assumption of two stimuli (i.e., $N=2$), there are exactly four states of conditioning corresponding to the four possible subsets of stimuli conditioned to the A_1 response. The complement of each subset is the set of elements conditioned to the A_2 response (in the present analysis we shall assume on every trial each stimulus is conditioned to exactly one response and that therefore there are no unconditioned stimuli on any trials). Representing the states of conditioning by the subset of elements conditioned to A_1 we then have the following notation for the four states: $\{s_1, s_2\}$, $\{s_1\}$, $\{s_2\}$, 0, where 0 designates the empty set.

To show how the assumptions of stimulus-sampling theory are used to derive a transition matrix in terms of these four states, we may draw the tree of possibilities with the probabilities for each branch shown when we begin in a typical state, let us say $\{s_2\}$.



By drawing trees for the other three states of conditioning to show what possibilities may arise at the end of the trial for each of the three when we start from that state, we may upon completing the trees, collect terms and obtain the following transition matrix:

	$\{s_1, s_2\}$	$\{s_1\}$	$\{s_2\}$	0
$\{s_1, s_2\}$	$1 - c(1-\pi)$	$c(1-\pi)^2$	$c\pi(1-\pi)$	0
$\{s_1\}$	$c\pi(1-\pi)$	$1 - 2c\pi(1-\pi)$	0	$c\pi(1-\pi)$
$\{s_2\}$	$c\pi^2$	0	$1 - c(\pi^2 + (1-\pi)^2)$	$c(1-\pi)^2$
0	0	$c\pi^2$	$c\pi(1-\pi)$	$1 - c\pi$

Casual inspection shows that this transition matrix is not the same as the one derived for the Bayesian models. On the other hand, if we let $c = 1$, we obtain the following special case of the stimulus-sampling model:

	$\{s_1, s_2\}$	$\{s_1\}$	$\{s_2\}$	0
$\{s_1, s_2\}$	π	$(1-\pi)^2$	$\pi(1-\pi)$	0
$\{s_1\}$	$\pi(1-\pi)$	$1 - 2\pi(1-\pi)$	0	$\pi(1-\pi)$
$\{s_2\}$	π^2	0	$1 - (\pi^2 + (1-\pi)^2)$	$(1-\pi)^2$
0	0	π^2	$\pi(1-\pi)$	$1 - \pi$

And it is immediately apparent that the entries in this matrix are the same as those for the Bayesian model. The identity between the two matrices also suggests, what should already have been apparent, that is, the formal isomorphism between the states of the Bayesian model and the states of the stimulus-sampling model. The following correspondence

$$\{h_1, h_{2'}\} \to \{s_1, s_2\}$$
$$\{h_1, h_2\} \to \{s_1\}$$
$$\{h_{1'}, h_{2'}\} \to \{s_2\}$$
$$\{h_{1'}, h_2\} \to 0$$

may be used to establish the isomorphism with the restriction that $c = 1$. Without this restriction the stimulus-sampling model is a slight generalization of the Bayesian one.

It may be remarked that the stimulus-sampling model, with c estimated from experimental data, does not appear to fit data very well (for some results in this connection, see Suppes and Atkinson [1960], Ch. 10). Of course, many Bayesians would almost be pleased that the stimulus-sampling model did not fit well for they could say "I would hardly expect the simple Bayesian model you have defined to provide any sort of decent fit to human

prediction data". The reply to this is straightforward. This same kind of Bayesian model may easily be extended to memories of finite length greater than 1, but immediately a common problem of the most essential sort for either the Bayesian or stimulus-sampling models arises. The problem is that we very quickly find ourselves in a combinatorial jungle out of which it is not easy to find a path. Consider, for example, the Bayesian model with finite memory of length 4 and again let us concentrate only on the pattern of reinforcement ignoring, although it is unrealistic, the pattern of preceding responses. For a finite memory of length 4 there will be 16 patterns of preceding reinforcements and thus a strategy will consist of a 16-tuple telling the Bayesian what to do when each of the 16 patterns is realized in the preceding 4 trials. This means there are $(2)^{2^4}$ or $(2)^{16}$ strategies to consider and thus this many states in the associated Markov process. The stimulus-sampling model with the additional parameter c has the same sort of difficulty. It is tedious but not impossible to obtain some results for models with this number of states. As the number of states increases, it rapidly becomes more difficult.

The generalized conditioning models applied to a variety of data in Suppes and Atkinson [1960], Suppes and Schlag–Rey [1962a] and Suppes and Schlag–Rey [1962b] may prove useful in examining in more detail the relationships between Bayesian and stimulus-sampling ideas. The essential idea of these models is to generalize the probability c of conditioning to let the probability of conditioning depend upon preceding responses and reinforcements. For the kind of application discussed particularly in Suppes and Schlag–Rey [1962a], one gets a formulation of conditioning models that is very similar to a kind of probabilistic Bayesian model with finite memory. It would seem to be primarily a choice of language and not of concepts as to how one prefers to describe these models. We described them as conditioning models but it is a simple matter to translate this description into Bayesian language. Some additional remarks about these models are made in the next section.

4. The structural problems besetting the theory of concept formation. In Section 2, I tried to make the point that any simple Bayesian approach to decisions, actions or choices, encounters considerable difficulty in explaining or predicting behavior of human subjects even in simple concept-identification experiments. I also tried to describe there some approaches that seemed promising from the standpoint of mathematical learning theory and, in particular, the version which originates with Estes and is ordinarily called stimulus-sampling theory. In order not to draw the distinction between

Bayesian models and stimulus-sampling models in too absolute a fashion, in the third section I tried to work out some of the formal similarities between the two approaches and in this case I chose for discussion a familiar paradigm in experimental psychology of recent years, namely, the two-choice situation with a noncontingent probabilistic schedule of reinforcement. In discussing various Bayesian models and the formally similar stimulus-sampling models that may be used to analyze the noncontingent case, I tried to sketch some of the combinatorial problems that quickly arise when more complicated and subtle models are considered.

To mention these combinatorial models first in connection with the noncontingent case is almost a mistake, for fairly simple stimulus-sampling models of a rather different sort than the kind considered in the preceding section give quite a good account of much data from noncontingent experiments. I have in mind the kind of pattern stimulus-sampling models first discussed in Estes [1959], and also discussed in Suppes and Atkinson [1960] Ch. 10, and Atkinson and Estes [1962]. By considering the standard pattern model of stimulus-sampling theory, it is possible to bypass some of the combinatorial problems I mentioned that arise for Bayesian models and the particular stimulus-sampling models corresponding to these Bayesian models.

The point of this section is to examine empirical situations, or simplified experimental situations roughly corresponding to the empirical situations in which it does not seem possible to avoid these combinatorial problems. It is a fundamental thesis of this paper that it is in dealing with situations in which new concepts must be formed that the standard formulations of the Bayesian approach are most inadequate.

To give the discussion some definiteness and concreteness, I shall primarily restrict myself to description of a class of experiments in which the problem facing the subject is to learn the grammar of a set of strings. It is to be emphasized, however, that in dealing with this grammatical example, I think of the problem of concept formation as being of a quite general nature. The difficulties besetting this example, particularly those of a combinatorial nature, apply equally well to any attempt to understand how humans learn to play well a complicated game like chess or make decisions rapidly when confronted with an incredibly wide choice of alternatives.

To fix our ideas quite specifically, let us consider initially the thirty-two strings of length five made up of 1's and 0's. From a formal standpoint we may define a grammar for this set of strings as a subset of the set of thirty-two strings. The number of such grammars is then $2^{32}-2$, where we exclude the universal and the empty grammar. Let us suppose that the subject is shown

the cards one at a time and is asked to classify them as codes or non-codes, where we think of a code as being a grammatical string and a non-code as being a non-grammatical string. The theoretical problem is now to describe how the subject proceeds to find the correct grammar. A simple Bayesian approach would be to attempt to describe a subject's a priori distribution on the $2^{32}-2$ possible grammars, and then to change this distribution as information is given to the subject concerning the classification of strings. It is just possible that for strings of length five something can be made of this Bayesian approach. For strings of length seven or eight or for anything approaching the complexity of chess, we must turn to the imposition of a considerable structure on the set of all possible grammars. It is, I would take it, the central problem of a theory of concept formation to provide such a structure and to state the laws by which organisms use the structure to solve the problem confronting them.

One way of approaching the problem of characterizing the structure of the space of all grammars is the following. The idea is to express any possible concepts for solving the problem of classification as a point in the space of properties associated with the stimulus material of the problems. A new concept is formed by moving to a new point in the property space. In these terms the theory of concept formation relevant to solving a given set of problems consists of two parts: first, characterizing the appropriate space of properties, and, secondly, characterizing the laws of motion in the space. In terms of the kind of formulation of stimulus-sampling theory considered in earlier sections, it may be thought that the phrase, "laws of motion", is too grandiose, and that what is described are simply the assumptions for sampling properties or stimuli. My reply to this possible objection has already been stated. The usual formulations of sampling assumptions neither assume nor impose any substantial structure on the set of stimuli (or concepts). The point of the present formulation is to impose such a structure. The space of properties is conceived as a multi-dimensional space with each dimension corresponding to a property. (Admittedly in many applications the space will consist of a finite set of points and thus will not satisfy the usual mathematical definition of a multi-dimensional space, but that is not a matter of serious concern here. I shall use the word "dimension" the way it is used in the psychological literature of concept formation and not in a mathematical sense.) It is only after a space is postulated (i.e., a set with a structure), that it is possible to talk about motion in the space. The concept of motion in an arbitrary set with no postulated structure is not well defined. On the other hand, it is precisely the imposition of structure that seems to be necessary to

bring some order and constraints to the discouragingly large number of possible concepts that may be considered in solving even a relatively simple problem. Once such a structure is imposed, laws of motion for the space, particularly when formulated as laws governing random walks, can be formulated.

To illustrate some of the possibilities for constructing the basic properties we may look at the problem when the set of strings is only of length two and, as before, at each position in the string there occurs either a 1 or a 0. According to the computations already indicated above it is immediately apparent that there are then 14 possible grammars for the set of four possible strings, excluding as before the universal grammar and the empty grammar. A simple ideographic space for this problem is the four-dimensional one, with one dimension for each card. The value on a given dimension is 1 if that string belongs to the hypothesis, and 0 otherwise. It is then trivial to represent any hypothesis as a point in this four-dimensional space. Such an ideographic space is not too unwieldly when the number of possible strings is small, but as has already been remarked, when this is the case the whole apparatus of a property space and the imposition of structure on this space is scarcely necessary. We may just as well use a straightforward stimulus-sampling or Bayesian model.

A more natural space of properties, which would generalize to longer strings, is the following. Dimension 1 characterizes the first position. The value 1 on the first dimension indicates that symbol 1 must occur in the first position of a string, and the value 2 that symbol 0 must occur in this position. The value 0 on this dimension indicates that either a 1 or a 0 may occur in the first position. The second dimension is defined similarly in terms of occurrence of symbols in the second position of a string. The third dimension is defined in terms of agreement or difference between occurrence of symbols in the two positions of the string. The value 1 in the third dimension is taken to indicate that the symbols occurring in the first and second positions of the string must be the same, the value 2 to indicate that the first and second positions of the string must be occupied by different symbols, and as before, the value 0 that the first and second positions may be occupied by the same or different symbols. We have selected the 0 value for all dimensions to indicate that this dimension is not intuitively relevant to the concepts, hypothesis, or grammar in question. The first thing of course to be noticed about this three-dimensional space is that there are a number of points that can be occupied by no concept that is nontrivial. Thus the concept to be represented by the coordinates (1, 2, 1) is the empty grammar, for it is not possible for a string to

have a 1 in the first position, a 0 in the second position, and yet to have the first and second positions occupied by the same symbol. Property spaces for other grammars or concepts connected with strings of this character indicate that this phenomenon is not easily eliminated. There does not seem to be a natural and simple way of defining orthogonal dimensions, but this does not seem to be an immediately crucial problem.

Still another way of looking at the space of properties is in terms of properties of a given string rather than of the grammar of the set of strings. In this case the grammar is represented by a certain subset in the space of properties rather than as a point. Corresponding to the space just constructed a space of this sort is easily described for the strings of length two, but I shall not go into details, because the present stage of our analysis of these problems, reinforced by some preliminary experimental evidence, indicates that this latter method is not the most desirable theoretical approach.

The space I did describe above for the strings of length two is deceptively simple. The extension of this same kind of description to strings of lengths greater than two soon becomes rather awkward if sufficient dimensions are required to locate with precision any grammar (or concept) in the set of all grammars. From experiments now being undertaken with Madeleine Schlag-Rey and some related experiments being conducted with elementary-school children in conjunction with Irene Rosenthal, it appears that for purposes of initial simplification of analysis we may in the case of strings of length three reduce the dimensions of the property space to a fairly small number, and lump the remaining unusual and not-likely-to-be-thought-of properties together. To give some rough indications of what we are finding, let me describe briefly the situation. In dealing with strings of length three, with the strings being built up from two symbols, there are $2^8 - 2 = 254$ possible nontrivial grammars. We have found, however, that about 80–85 percent of the grammars conjectured by subjects may be classified under six main property headings and so we have restricted the analysis to these six properties together with a catch-all seventh category in which we place the remainder. The main point of our investigation at the present time is to find out to what extent the behavior of subjects in selecting and rejecting grammars (or more generally concepts or hypotheses) may be accounted for in terms of the application of stimulus-sampling models to the seven properties. To use the physical language mentioned earlier we are attempting to characterize the motions of the subjects' changes in concepts or hypotheses in terms of random walks with respect to the most salient properties, as for example, the occurrence of a 0 or 1 in one of the three positions, or the occurrence of a

matching pair in the first and second, the second and third, or first and third positions.

It is too early yet to decide whether or not this particular approach to concept formation will prove to be a fruitful one. Before concluding, I do want to indicate how the generalized conditioning models studied in Suppes and Schlag–Rey [1962a], which were mentioned earlier, have a bearing on finding the properties which are most salient for subjects in structuring their approach to the solution of a problem. The experiment analyzed was one with a probabilistic reinforcement schedule in which the reinforcement in a two-choice situation, on a given trial, depended or was contingent upon the subject's own preceding two responses. We were particularly concerned to analyze the experimental data to find the nature of the patterns to which subjects seem most likely to condition their responses.

Basic data examined in the experiment were the conditional probabilities of an A_1 response given the reinforcements and responses of the two preceding trials. Ten different models, each postulating that the conditioning of the responses depended on a different pattern, were considered. In Class I of the models the sequential dependence or conditioning was defined in terms of the two physical sides 1 and 2 of the key and light apparatus (of course, for some subjects side 1 was on the left side and for some subjects on the right). The point is that the conditioning parameters in Class I were defined in terms of the side. The five special cases considered in this Class were defined by restricting the dependency of $A_{1,n+1}$ to: (a) the response and reinforcement that occurred on trial n; (b) the two preceding reinforcements; (c) the two preceding responses; (d) the two preceding reinforcements and the immediately preceding response; (e) the two preceding responses and the immediately preceding reinforcement.

In Class II the conditioning parameters were defined, not in terms of the sides 1 and 2, but in terms of successful and unsuccessful responses, rewarding and punishing reinforcements, repetition or alternation responses, etc. In particular the five special cases were defined by: (a) the reinforcement on trial n was punishing or rewarding; (b) the reinforcements on trials $n-1$ and n were punishing or rewarding; (c) the reinforcement on trial n was punishing or rewarding, and the response of trial n indicated anticipation of a repeating or alternating reinforcing event; (d) the reinforcement on trial n was punishing or rewarding, and the response on trial n was a repetition or alternation of the response on trial $n-1$; (e) the reinforcements on trial $n-1$ and n were punishing or rewarding, and the response on trial n was a repetition or alternation of the response on trial $n-1$.

From the standpoint of the present paper these ten models provided an opportunity for gathering information on the kind of structure subjects tend to impose in such a probabilistic situation. The observed transition probabilities and the goodness-of-fit tests for the ten models of Class I and Class II are given as Table 2 of Suppes and Schlag-Rey [1962a], and will not be reproduced here. The most important observation about the results of the goodness-of-fit tests, however, is that with the same net number of degrees of freedom the fits of Class II models were uniformly better than Class I. In addition, the assumption that the conditioning can be explained in terms of the last reinforcement's being punishing or rewarding yields a better fit than did any of the Class I assumptions with four parameters. The uniformly better results of the Class II models in comparison with the Class I models supports the hypothesis that subjects are in many cases more likely to sample patterns of stimuli defined in terms of complex relational properties than in terms of relatively concrete single events. Detailed information about the relative saliency of such relation-defined patterns is one of the most important things needed to move ahead with an empirically adequate theory of concept formation.

The ideas about concept formation set forth in this paper are meant to be suggestive rather than definitive. I do hope, however, that the various kinds of examples considered present adequate evidence for maintaining that any theory of complex problem solving cannot go far simply on the basis of Bayesian decision notions of information processing. The core of the problem is that of developing an adequate psychological theory to describe, analyze and predict the structure imposed by organisms on the bewildering complexities of possible alternatives facing them. The simple concept of an a priori distribution over these alternatives is by no means sufficient and does little toward offering a solution of any complex problem.

Moreover, understanding the structures actually used is important not only for an adequate descriptive theory of behavior but also for any normative theory intended to be applicable to human beings with finite powers of memory and computation. As the standard literature of inductive logic comes to grips with more realistic problems the overwhelming combinatorial possibilities that arise in any complex problem will make the need for higher-order structural assumptions self-evident.

References

ATKINSON, R. C. and W. K. ESTES, 1963, *Stimulus sampling theory*, in: Handbook of

Mathematical Psychology, vol. 2, eds. R. R. Bush, R. D. Luce and E. Galanter (John Wiley and Sons, Inc., New York) pp. 121–268

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

BOWER, G. H. and T. TRABASSO, 1964, *Concept identification*, in: Studies in Mathematical Psychology, ed. R. C. Atkinson (Stanford University Press, Stanford, California)

BOURNE, L. E. and R. RESTLE, 1959, *Mathematical theory of concept identification*, Psychological Review, vol. 66, pp. 278–296

ESTES, W. K., 1959, *Component and pattern models with Markovian interpretations*, in: Studies in Mathematical Learning Theory, eds. R. R. Bush and W. K. Estes (Stanford University Press, Stanford, California) pp. 9–52

SAVAGE, L. J., 1954, *Foundations of statistics* (John Wiley and Sons, Inc., New York)

SUPPES, P. and R. C. ATKINSON, 1960, *Markov models for multiperson interactions* (Stanford University Press, Stanford, California)

SUPPES, P. and R. GINSBERG, 1962a, *Application of a stimulus sampling model to children's concept formation with and without an overt correction response*, Journal of Experimental Psychology, vol. 63, pp. 330–336

SUPPES, P. and R. GINSBERG, 1962b, *Experimental studies of mathematical concept formation in young children*, Science Education, vol. 46, pp. 230–240

SUPPES, P. and R. GINSBERG, 1963, *A fundamental property of all-or-none models, binomial distribution of responses prior to conditioning, with application to concept formation in children*, Psychological Review, vol. 70, pp. 139–161

SUPPES, P. and M. SCHLAG-REY, 1962a, *Test of some learning models for double contingent reinforcements*, Psychological Reports, vol. 10, pp. 259–268

SUPPES, P. and M. SCHLAG-REY, 1962b, *Analysis of social conformity in terms of generalized conditioning models*, in: Mathematical Methods in Small Group Processes, eds. J. Criswell, H. Solomon and P. Suppes (Stanford University Press, Stanford, California) pp. 334–361