LEARNING AND APPLYING CATEGORY KNOWLEDGE IN UNSUPERVISED DOMAINS

John P. Clapper
Gordon H. Bower

I. Introduction

In order to behave intelligently, people need internal models of their environment and how their actions will modify it (see Craik, 1943; Gentner & Stevens, 1983; Johnson-Laird, 1983). In this article, we focus on how people discover regularities among the objects, events, and situations they encounter, and how they create general categories to capture these consistent patterns. By partitioning their experiences into distinct categories, people can build up a collection of internal models that apply to many similar objects or events, allowing them to use their past experience to interpret and respond adaptively to the current situation. We are also concerned with the way that such categories, once formed, affect how further instances are encoded into memory and how such encodings serve as a basis for discovering further subcategories within a given domain. Much previous research indicates that learning to recognize recurrent patterns (categories) within a stimulus domain improves the efficiency of encoding and representing specific instances (e.g., Chase & Simon, 1973; de Groot, 1965, 1966); this, in turn, should facilitate the discovery of new, more specific subcategories within that domain.

Our analysis presupposes a broad definition of categories that focuses on their role in allowing a knowledgeable subject to predict the features of
specific instances. Here, a category is defined as any collection of features or components that occur together with relatively high consistency across different contexts (instances). Such correlation of elements or constituent features provides the learner with predictive power: Given that enough features are observed in a specific instance to match it up with the appropriate category, the category can be used to predict or imply the presence of other features not directly observed. This broad definition admits many types of "categories" that are not always studied or thought of as such, including recurring temporal patterns, chess board configurations, mental models of standard electrical circuits, and so on, as well as familiar "natural kinds" such as cats, birds, or medical diseases. While there may be real differences between different types of categories, e.g., between those with a clear object–property structure as opposed to, say, recurring temporal patterns in a musical score, those differences are beyond the scope of the present analysis.

II. Theoretical Issues

A. Induction of Categories from Instances

A large experimental literature has accumulated regarding the acquisition of concepts, much of it following in the tradition of studies by Bruner, Goodnow, and Austin (1956; see, e.g., Millward, 1971, for a review). Most of that laboratory research focused on supervised learning of categories, in which an external tutor informs the subjects what classifications are to be learned (i.e., category labels are provided by the tutor) and provides feedback relative to a specific criterion for the current learning task. By contrast, many categories that people acquire in real life are learned in untutored, observational conditions. Such unsupervised learning occurs in the absence of predefined categories and without feedback from any tutor. For example, much of what we know about the perceptual properties and behavior of physical objects, social interactions, linguistic classes and rules, and everyday tasks and procedures is probably learned in this manner (Billman & Heit, 1988). Any learning of a pioneer in a novel environment is unsupervised. For example, botanists classifying new plants from a newly discovered island, geologists classifying rocks from a new planet, or medical pathologists inspecting histological sections of tissues infected with various diseases have no one to tell them how to group the specimens in different subclasses. Pioneers must create their own groupings that are sensitive to the salient regularities detectable in the domain.

Curiously, little psychological research or theorizing has been devoted to the topic of unsupervised learning. This paucity of research is surprising in light of the pervasiveness of unsupervised learning in everyday life. More to the point, unsupervised learning seems to involve somewhat different principles from those that characterize concept learning with a tutor. For example, in an unsupervised task the learner must decide whether, and how, to create new concepts to describe stimuli that fit poorly into existing categories. This issue does not arise in supervised classification tasks because the tutor essentially tells the learner when to set up a new category and assign a given exemplar to it. In addition, unsupervised tasks provide opportunities to study people's incidental learning of concepts; by studying when certain concepts spontaneously "pop out" of the learner's stream of experience, we may discover the kinds of regularities that people's inductive machinery is designed to detect naturally, in contrast to the regularities whose learning requires a tutor's feedback.

B. Representation of Categories in Memory

For centuries philosophers have debated the mental representation of concepts. One view proposes that learners abstract summary representations of categories (e.g., prototypes or schemas; see Kant, 1787/1963; Posner & Keele, 1968, 1970; Rumelhart & Ortony, 1977); a contrasting view proposes that people merely store in memory collections of instances from which generalizations about the category can be computed as they are needed (e.g., Hintzman, 1986; Hume, 1748/1960; Medin & Schaffer, 1978). This debate has proven difficult to resolve because the range of inferences derivable from unrestricted computation over a collection of stored instances is greater than that provided by summary statistics, such as mean and variance, computed from these instances. In other words, any generalization or inference about a category (e.g., boundary conditions for membership) that could be stored in a summary representation could also be computed during testing from memories of specific instances. For this reason, it is difficult to obtain strong evidence for summary models merely by examining the pattern of classification responses people give to a particular set of test stimuli.

Although instance storage theories can easily mimic the inferential power of summary representations, the two positions differ in their assumptions about (1) how knowledge of categories, subcategories, and individual exemplars is organized in memory and (2) people's ability to use their experience with a category to improve their interpretation, analysis, and encoding of specific novel cases. These differences mainly influence measures of processing or encoding in various tasks rather than the semantic content of subjects' inferences. For example, different theories of
memory organization make different predictions about which factors affect the speed and accuracy of verifying facts from memory, even though the theories may agree on what features subjects would attribute to a given category or instance.

In the experiments to be described, we investigated how knowledge of a category affects the learning of new instances. Many of the patterns of learning and memory organization observed in our experiments appear incompatible with current formulations of instance theories. We do not take such results as implying that people never retain specific instances or make inferences based on them; rather, we interpret results as showing that under appropriate conditions people are also quite capable of learning and applying generalizations about members of a category.

C. Category-Based Processing of Instances

Considerable research suggests that concept models (e.g., spatial models, temporal scripts) play a key role in how people process instances. This role is illustrated in studies comparing experts to novices in processing problems in their domain. These studies show that experts are able to represent information about their domains more efficiently than novices, resulting in much improved memory and problem-solving performance (e.g., Chase & Simon, 1973; de Groot, 1965, 1966). This efficient compacting of information by experts is seen in domains as diverse as chess, electronic circuitry, baseball, and culturally specific scripts. Experts' advantage rests on their stockpile of categories for recognizing recurrent situations in their domains, so that such situations can be represented and reasoned about with familiar ideas.

Interestingly, the experimental literature on concept learning has largely ignored the way that concepts, once formed, alter the processing of later instances. The traditional research agenda has been driven by another issue, i.e., how to characterize the essentially "bottom-up" process by which people acquire generalizations about categories from descriptions of specific instances (e.g., Hunt, 1962; Millward, 1971). This learning orientation may be contrasted to the "top-down" process by which people use their category knowledge to guide their processing of instances. Consistent with this traditional emphasis, concepts have tended to be regarded as decision rules for classifying new stimuli, but not as active processing structures that determine how particular stimuli are represented and acted on.

Much of the existing research on how general models of categories affect the interpretation and processing of specific situations has used social-ethnic stereotypes (e.g., Srull & Wyer, 1989), personality stereotypes (e.g., Cantor & Mischel, 1979), or situational scripts for routine activities (e.g., the restaurant script of Schank & Abelson, 1977). An example of the latter type of experiment is one by Bower, Black, and Turner (1979), which examined how subjects' memory for text statements describing a routine activity varied with whether a given event was predictable or deviated from the script. At best, however, such experiments yield only imprecise measures of processing; their obvious imprecision arises from the experimenter's ignorance of the structure, properties, and training history of these familiar concepts for each subject. For valid generalizations, we prefer to investigate the way subjects learn and use artificial concepts that have been designed to precise laboratory specifications.

III. A Model of Unsupervised Learning

To guide our research, we have developed a tentative model describing how people might learn categories in unsupervised environments and use them to guide the encoding of specific instances. The learners in this model are assumed to be engaged in unguided exploration of a given domain of objects, i.e., learning is unsupervised and learners are simply attending to the features of individual objects without explicitly searching for categories among them. Importantly, human learners have a limited attentional capacity with which to carry out such exploration. This capacity limitation presents them with the problem of selecting appropriate features of their environment to attend to and record into memory. When the environment provides direct reinforcement (e.g., a tutor's feedback in a supervised classification task), subjects will learn to attend to features that are correlated with this reinforcement. In the absence of such explicit consequences, we assume that people use heuristic strategies for allocating attention that help them (1) encode instance representations with the greatest efficiency, given their limited attentional capacity, and (2) maximize the likelihood of discovering useful patterns or regularities, i.e., new subcategories, without explicitly searching through the space of possible categories within a given domain. We hypothesize that such heuristics require that subjects use existing categories to evaluate new stimuli and distinguish informative from uninformative features, and then selectively attend to the informative features. In the present treatment, we are particularly concerned with statistical, inductive determinants of informativeness, i.e., the relative likelihoods of different features within a category. We do not dispute that other factors may influence perceived informativeness in addition to the inductive criteria that we emphasize, e.g., theoreti-
contrasted. In practice, an experimenter will usually define a set of canonical attributes by which a stimulus set is generated and/or described, with each instance from that set represented as a vector of specific values. Within such a stimulus set, different subcategories are distinguished by clusters of correlated (consistently co-occurring) attribute values. For example, a geneticist might describe a collection of fruitflies in terms of several attributes such as size, wing shape, eye color, and so on. If it was then noticed that most individuals with long wings were also large and had red eyes, whereas those with short wings were medium-sized and had white eyes, these patterns of co-occurrences would form an inductive basis for distinguishing two different families or categories within that population. Importantly, this characterization of categories does not imply that the interfeature correlations must be perfect; in principle, a category would have positive utility as long as at least some features of instances could be predicted with greater-than-baseline reliability. Thus, the present characterization of categories admits “fuzzy” categories with probabilistic features and does not require categories to be defined by necessary and sufficient features. Moreover, not all the attributes of an instance would necessarily be highly correlated with its category; within any category, different attributes will take on predominate values with different reliabilities across different instances.

We assume that a learner’s knowledge of a specific category is represented in long-term memory as a schema that specifies a set of attributes in terms of which instances will be described, and specific values of each. Norms for each attribute are represented by a collection of strengths of association between the category and each value of the attribute. These strengths represent the relative expectedness or availability from memory of each value of each attribute, e.g., their frequency and recency among previous category members.

B. STEP 1: CATEGORIZE THE INSTANCE

A basic tenet of our approach (see Fig. 1) is that a stimulus event evokes its own frame of reference, i.e., it is automatically classified in the best-fitting category available from memory. This category then provides a familiar framework within which the stimulus can be further interpreted and reasoned about. A stimulus is categorized by matching a sample of its features (i.e., specific values of attributes) to the attribute norms for each candidate category, and then selecting the best match. An instance will be assigned to that category with which it shares the largest proportion of highly expected values and the fewest surprising or unusual values.
C. **Step 2: Evaluate the Instance**

With respect to the norms of a particular reference category, the features of a given stimulus will vary in how typical or expected they appear. This typicality in turn determines which features are considered most informative for describing that particular instance. Although several definitions of "informative" are plausible, all capture the intuition that the informativeness of a stimulus value increases with its unpredictability or surprisal in a given context. Importantly, this principle implies that consistent, highly expected values of an attribute (referred to as defaults) will be considered uninformative, whereas features that are unusual or not specified in advance by the schema will be judged as informative.

A simple way to conceptualize a value's informativeness is in terms of its strength (availability) in the attribute norm for a given category. The greater a value's strength, relative to alternative values of the same attribute, the more expected and less informative it should appear. Assuming that a value's strength increases with its relative frequency within a category, the strength view is equivalent to equating a value's informativeness with its improbability of occurrence within the category.

This view of informativeness is broadly consistent with a rational encoding strategy for an "ideal learner," and with intuitions about the kinds of events that people find interesting and to which they pay attention. An efficient learning mechanism should attempt to maximize the new (previously unknown) information it acquires about a stimulus within the encoding capacity available for a given task. Thus, an ideal learner should avoid expending limited resources recording facts about an exemplar that are already predictable from categorical knowledge; rather, the optimal strategy is to focus on features that are unpredictable, surprising, and informative.

Just as people tend to focus on informative features when recording a given experience, so do they also focus on similarly distinctive, informative properties when communicating to others. This injunction is embodied in one of Grice's "maxims of conversation," namely, that speakers should be informative and not convey known, redundant information to listeners (Grice, 1975). For example, uninteresting truisms are not normally uttered in conversation; rather, people abide by the rule of describing objects, situations, and events in terms of their more distinguishing or informative properties. Thus, people might describe their car as a "blue Chevy" but not as a "Chevy with four wheels"; they refer to penguins as "flightless birds" but not to dogs as "flightless mammals," although both statements are equally accurate. When describing criminal suspects, police bulletins and news programs highlight any unusual features the suspect might possess, such as scars and tattoos, rather than features the suspect shares with the general population.

D. **Step 3: Encode the Instance**

After categorizing the instance and assigning informativeness to each of its attribute values, the next step is to record the instance into memory. The features of an instance compete for fixed attentional or encoding resources, which are assumed to be distributed among the features so as to maximize the total informativeness of the features encoded. The model assumes that the resources allocated to a given attribute value are proportional to its informativeness relative to that of the other attributes of the stimulus. This rule ensures that the learner encodes the maximum of distinguishing information about an instance given the attentional resources available to process it.

The episodic memory representation of the instance that results from this encoding process can be thought of as a set or vector of features, each with a specific strength of association to that instance. A feature's strength in this record depends on how much attention it received at encoding, which depended in turn on its informativeness. The instance's categorization at the time of encoding is presumed also to be stored with that instance in memory.

E. **Step 4: Update Category Norms**

The model assumes that people incrementally update their norms for the activated concept after each presented instance. Two cases are distinguished according to whether the current instance is adequately covered by a previous category or, due to its novelty, requires the creation of a new category.

1. **Assimilation to a Previous Category**

Normally, instances are assimilated to the category used to evaluate and encode them. The attribute norms of this category are adjusted by increasing the strength of each observed value in proportion to the amount of attention it received during encoding. As the same value of an attribute is repeated over a series of instances, it becomes less informative and learners should pay progressively less attention to it. This process is analogous to habituation to a constant stimulus repeated within a particular context, except that the context in this case is given by category membership. Due to this habituation of the constant features, more attentional resources are left over to process the remaining, unpredictable features of each instance.
2. Create a New Category before Assimilating

Learners are assumed to create new categories in response to the failure of old ones. Specifically, in exploring a domain, learners use a “surprise heuristic” to indicate when they should invent a new category. According to this heuristic, when an instance contains sufficient features that are surprising (highly informative) with respect to its assigned category, a new category is created to describe the unusual stimulus. This strategy for creating new categories is similar to the “failure-based generalization” of Schank (1982). By creating new categories only when an instance violates prior norms, subjects can learn categories in a domain without explicitly searching through the entire space of possible categories within that domain, i.e., keeping track of all possible feature correlations. For complex domains characterized by vast numbers of possible categories, such an explicit search strategy might be unrealistic for human learners (but see Billman & Heit, 1988, for a different approach to solving this search problem).

The model assumes that if a new category is triggered by an unusual instance, then that instance will be assimilated to the new category; thus, the unusual instance will not affect norms for the prior concept from which it deviated. This segregation principle allows people to accommodate highly unusual instances without discarding beliefs that have proven generally useful and reliable in the past. To illustrate, if zoologists discovered an unusual elephant that had thick fur and no tail, they probably would not abandon their belief that elephants are generally hairless with tails instead, they would assume that they had discovered a new subspecies of elephant, closely related to, but distinct from, the familiar species.

In the model, the schema for the new category is created by modifying the schema for the “source” category (that to which the instance was originally assigned) in order to describe the deviant instance. In doing this, we assume that learners believe that all their norms about the source category that are not specifically violated by the triggering instance can be transferred to the new category created around that unusual instance. New norms are created only for those attributes whose exceptional values triggered the formation of the new category. To return to our example of the furry, tailless elephant, in creating a new category around this stimulus, the model would transfer all its prior beliefs about elephants to the new category (e.g., that they are plant eaters, have trunks and lungs, etc.), except those relating to the “fur” and “tail” attributes. New norms, based on the triggering instance, will be created for these unusual attributes; prior norms concerning these attributes for ordinary elephants would not apply to this new subspecies. By conforming to this transfer rule, learners need to make the fewest possible changes to their existing taxonomy to handle the deviant instance. Thus, existing knowledge is exploited to the fullest to conserve computational resources in forming the new concept.

F. Step 5: Retrieving Features from Instance Memories

When people attempt to remember the features of an instance, limited retrieval resources (e.g., spreading activation) are divided among the features in its underlying memory representation. The activation received by each feature increases with its strength divided by the combined strength of all features of that instance. This rule implies that the more features that are strongly associated with an instance, the more difficult it should be to retrieve any particular one. This fact has received extensive empirical validation in analogous memory experiments: the more independent facts that people learn about a particular topic or item, the more time they require to verify any one of them from memory (see Anderson, 1976, 1983, for reviews of this research). This phenomenon is known as the fan effect or as associative interference.

The model’s assumptions about encoding and retrieval imply differences in the way predictable vs. unpredictable features of an instance are remembered. Because of their low informativeness, the highly predictable features of an instance will be only weakly associated, if at all, to the instance. As a first approximation, we will simply assume that subjects omit these category defaults from their memory representations of specific instances. Rather, the default values would merely be noted as properties of the general category, and hence inferable for specific instances by property inheritance. In such a memory organization, the default properties of an instance would be effectively segregated from its distinctive or variable features. The instance with its distinctive features would be recorded as a “subnode” in memory pointing to the category node with its associated defaults (see Fig. 2). As a result, when retrieving the fact that an instance has a specific distinctive feature, the system avoids associative interference (the fan effect) from the category defaults. Besides economizing on learning and storage that results from the encoding process, this “subnoding” maneuver confers a major advantage on this memory organization for later information retrieval. The memory organization helps solve the so-called paradox of interference, which is that experts with vast domain knowledge do not suffer the massive slowdown in retrieval that interference principles alone would have expected (Smith, Adams, & Schorr, 1978). The subnoding solution is similar to earlier solutions of the paradox proposed by Reder and Ross (1983) and Anderson (1983).
IV. Comparison to Alternative Approaches

The model we have proposed differs in several respects from previous models of unsupervised learning. One advantage of the present model is that it makes explicit the role of generic concepts in the interpretation, analysis, and recording of novel cases; in turn, the model shows how the processing of specific instances affects the learning of category level expectations. Most previous models of category formation are strictly "bottom-up" in the sense that they specify how instance information is used to form concepts but not how the concepts in turn determine the encoding and representation of further instances. By exploring these issues experimentally, we hope to shed light on how concepts economize the processing of later exemplars.

Most previous models of concept learning were formulated to deal primarily with the classification of instances into categories, and not with the problem of storing those instances in memory for later reproduction. Consequently, they assume that learners become more likely to attend to attributes whose values consistently co-occur across category members (i.e., that are diagnostic of category membership). While this process is acceptable for partitioning a stimulus set into categories at one level of specificity, it is not adequate for learning and retrieving descriptions of specific instances or for building hierarchies of categories and subcategories at multiple levels of specificity. A classification model that increasingly focused attention on category diagnostic features would learn progressively less about the distinguishing features of specific instances. Similarly, a learning process that focuses solely on known category defaults would be completely blocked or very slow in learning specific subcategories that might be differentiated within more general categories. For example, once having learned to differentiate oak trees from maple trees, people operating under this limitation would be unlikely to attend to subtler properties that differentiate among subspecies of oaks because they would be focusing instead only on features common to all oaks. Such a focus contrasts with more naturalistic learning, in which people consider known categories as "background" and proceed to focus on subtler distinctions among instances that might form a basis for learning more differentiated categories.

V. Experiments

A. EXPERIMENT 1: ATTRIBUTE LISTING TRACKS

UNSUPERVISED LEARNING

In a first experiment, we explored a new task designed to investigate the evolution of subjects' category level norms as they examined successive instances from a single category. The objective of this task was to provide a trial-by-trial index of subjects' evolving beliefs about the informativeness of each attribute and its specific values. According to the model, subjects should learn to discriminate among the features of instances according to their informativeness within the reference category. Category defaults should be considered uninformative and receive low priority, while exceptional or highly variable features should receive high priority. These biases should develop gradually as subjects accumulate experience with instances of a given category. If an index could be found for the informativeness of each feature in a series of training instances, this index could be used to trace "learning curves" for norms about the experimental categories. By studying the properties of such learning curves and how they are affected by task structure and stimulus design, much fundamental knowledge could be acquired about unsupervised learning.

In this experiment, subjects were shown a series of training instances from a single category and were asked to list the distinguishing features of each. The distinguishing features were portrayed for the subject as those characteristics that would be most helpful in discriminating that instance from others of the same general type on a multiple-choice recognition test. The stimuli were line drawings of fictitious insects (see Fig. 3). The insects were composed of a consistent base structure, consisting of parts such as
instances, in which all defaults were present as described above. After these initial training trials, however, subjects would occasionally see a deviant instance in which a particular default was violated (e.g., if all the standard bugs had wings, one of the deviant bugs would be wingless). Out of a total of 40 instances presented over the course of the experiment, subjects saw two such deviant instances, each of which violated a different default.

The dependent measure of interest was the proportion of subjects that listed each of the nine experimental attributes on each trial. The probability of listing the presented value of an attribute should depend on how informative subjects consider it to be for discriminating that instance from other category members. Thus, such listing provides an index of changes in subjects’ learning about that attribute at each point in the training sequence.

Turning to the results, as expected subjects’ reporting of default values declined rapidly over the first few trials, from about 54% on the first trial to 10% on the fifth, where it remained thereafter (see Fig. 4). The linear component of this decreasing trend was statistically significant at the .01 level \( t(24) = 5.69 \), and the quadratic component was marginally significant \( t(24) = 1.79, p < .10 \). Subjects’ reporting of variable attributes averaged about 75% and was fairly constant over trials. By the end of the training phase, the variable attributes were being reported nearly 65% more than the default attributes, a highly significant difference \( p < .01 \).

This pattern of results indicates that subjects rapidly learned that the presence of defaults could be taken for granted and that only the variable features provided differentiating information about each instance.
A second result of interest was that subjects were very likely to notice and report the absence of default features in the two deviant instances. The increase in listing, from 10% on the preceding trial to 72% on the deviant trial (marked with an asterisk in Fig. 4), was highly significant \( t(24) = 8.40, p < .01 \). Listing of the default value dropped significantly on the following, normal, trial \( t(24) = 5.52, p < .01 \), but for several trials remained higher than it had been previously. This result suggests that the missing default caused subjects to temporarily “dishabitate” to that attribute, much as an unexpected change in a stimulus produces an orienting reflex and temporarily releases previous habituation to that stimulus.

These results indicate that the attribute listing task is sensitive to manipulations of feature informativeness and that the patterns of attribute listing are consistent with the model of unsupervised learning. The findings suggest that attribute listing is a useful method for tracing learning curves for subjects’ attribute norms as successive instances are assimilated to a single category. Thus, the method should prove useful for investigating many variables that influence unsupervised learning.

B. EXPERIMENT 2: LEARNING TWO BLOCKED CATEGORIES

The next experiment proposes extending the attribute listing procedure to a situation in which subjects would learn two contrasting categories in an unsupervised environment. This would require that subjects learn the category level discrimination built into the stimulus set and that they reflect this learning in their patterns of attribute listings. That is, subjects should learn to selectively report the values of attributes that are variable within each category, while omitting values shared by instances within a category (but that differ between the two categories). Such a response pattern would indicate that subjects had learned to evaluate a value’s informativeness within the specific category to which the current instance belongs rather than evaluating its informativeness across the stimulus set as a whole. Hopefully, the results would show separate learning curves for the attribute norms of each category. If successful, the method would enable investigation of variables that facilitate or interfere with the discovery of distinct categories, thus permitting evaluation of different models of how such discoveries arise.

The stimuli were similar to those used in Experiment 1, i.e., line drawings of fictitious insects (see Fig. 3) that varied along dimensions such as wing type, eye color, length of legs, and so on. Two distinct categories were defined by collections of correlated values on several attributes. For example, for a given subject all of the insects in the A category might have wide wings, a fat body, fuzzy antennae, large pincher mouthparts, and black eyes, whereas members of the B category would have opposite values on each of these five attributes. A total of eight attributes in this stimulus set could be varied to create different instances. Of these eight, five were assigned a consistent default value within each category, whereas the remaining three were free to vary across instances. Within categories, each variable attribute had two values (a different two for each category) that occurred equally often across different instances and varied independently of each other. Two different stimulus sets were created and shown to different groups of subjects to ensure that the assignment of attributes to the variable vs. default conditions was properly balanced.

The procedure and instructions to the subjects were similar to those of Experiment 1. The stimuli were presented in booklets with one insect per page and a space at the bottom for subjects to write their feature lists. The instances were presented blocked by category. Sixteen instances of the A category were presented first, followed by 16 instances from the B category. Such blocking should increase the probability that subjects would create two distinct categories rather than assimilating both A and B instances to a single omnibus concept. Following the two same-category blocks, a final block of eight trials was presented in which instances of the two categories were intermixed in a random sequence. This mixed block was included to check whether the discrimination learned during the blocked trials would be maintained when instances of the two categories were presented in random order.

Consistent with the results of Experiment 1, subjects learned the defaults of the A category as they examined the first several instances, gradually reducing their listing of the default features of category A. Starting from a high rate of listing of 58% on the first trial, listing of default values gradually decreased over trials until it reached about 16% on trial 10, where it remained until the first B instance was presented (see top panel of Fig. 5). This decreasing trend in subjects’ listings of A defaults was highly significant \( t(15) = 4.29, p < .001 \). By contrast, the A variables (bottom panel of Fig. 5) were listed with an average rate of 94%, significantly exceeding that of the defaults \( t(15) = 12.87, p < .001 \). Upon encountering the first instance of the B category, subjects dramatically increased their listing of the contrasting defaults for that new category (Fig. 5, top panel). The 55% listing in listings was highly significant \( t(15) = 7.00, p < .001 \). Over the next several trials, however, subjects gradually reduced their reporting of these newly constant defaults and reverted to a strategy of listing mainly the variable features of each instance. This decreasing trend was significant at the .001 level \( t(15) = 4.39 \). These listing patterns reveal orderly learning curves for the acquisition of the two concepts.
On that account, the gradual decrease in listings in both blocks would reveal merely localized habituation within a single category due to lengthy "runs" of similar instances. In other words, during the series of B instances, subjects would have gradually modified their attribute norms so that the earlier A defaults were now overshadowed by the more recent B defaults. Were this the correct account, subjects should have dramatically increased their listing of these attributes when they encountered the first A instance during the mixed block (since the A defaults would now be considered exceptional). Moreover, by this alternate account, if default attributes began to appear highly variable throughout the mixed trials, then they should have been listed at the same frequency as the variables in earlier trials. The fact that subjects continued to list the A category (and B category) defaults with low frequency during the mixed block indicates that they had acquired two distinct, stable concepts during the blocked trials.

The results of this experiment demonstrate that the attribute listing task can trace the discovery and learning of distinct categories in an unsupervised task. The pattern of learning observed in this experiment is consistent with the predictions of the information-processing model described earlier. This task may therefore have considerable potential for exploring variables that affect category discovery and modification, allowing tests of the basic model.

C. Experiment 2A: Learning two categories with mixed sequences

Experiment 2 demonstrated that subjects could learn to discriminate categories based on contrasting defaults without supervision from an external tutor, and that the attribute listing task could be used to index this learning. A possible criticism of that experiment, however, is that the training instances were artificially sequenced by categories to maximize the probability of successful discrimination learning. Perhaps such learning would not have occurred with a less contrived, random sequence of A and B exemplars. In defense of Experiment 2, it was explicitly designed to demonstrate that subjects could detect contrasting categories without supervision when the exemplars were optimally sequenced and that such learning would be reflected in their attribute listings. Experiment 2 did not explore the boundary conditions under which such learning would be possible. We wanted to demonstrate in the next experiment (2A) that unsupervised learning could occur even with intermixed training sequences.

The information-processing model sketched in the Introduction predicts...
that subjects could learn to discriminate contrast categories in a randomly intermixed sequence, but that in some cases this learning would be considerably impaired relative to a blocked-sequence condition. Such interference is expected on the model's assumption that new categories are created in response to expectation failures (surprise); for the defaults of a B instance to appear surprising, however, they must violate strong default expectations already acquired for category A. It is easy to understand how such surprise would arise in Experiment 2, since strong expectations would be acquired during the A block before the first B instance was encountered. However, when a randomized training sequence is used, only a small number of A instances (perhaps one or two) would be encountered prior to the first B instance, providing little opportunity to learn strong A defaults. According to our theory, violations of weak expectations are less informative (surprising) than violations of strong ones; thus, the new value may not appear sufficiently surprising to trigger the creation of a new category. Instead, the two types of instances might be assimilated together into a single, encompassing category that summarized the stimulus set as a whole. In such a case, the default values of each category would be encoded as alternative values of variable attributes, i.e., since the correlations among the default values would not be captured by having separate categories, subjects would be unaware of them and would consider the values uncorrelated. Once the categories were initially aggregated together in this manner, it might be difficult to later "unlearn" this overgeneralized framework and correctly discriminate the two categories. Indeed, an analogous effect on supervised category learning was demonstrated by Holstein and Premack (1965), who found that an initial period of random feedback substantially retarded learning of a simple classification of the same stimuli.

Experiment 2A was similar in most respects to Experiment 2 except that the stimuli were presented in an intermixed sequence rather than blocked by category. A total of 48 instances was presented, designed according to the same specifications as in Experiment 2. The order of instances was randomized, except that runs longer than three instances from the same category were disallowed. As before, five attributes of the stimuli had correlated values that served as defaults for the two categories, while the remaining three varied independently across different instances. Two different stimulus sets were constructed, each with a different assignment of attributes to default vs. variable status, to ensure that any attribute-specific effects would be properly counterbalanced. In addition, two other stimulus sets were constructed in which all eight attributes varied independently. These fully variable "control sets" allowed us to compare listing performance in structured vs. unstructured stimulus sets, and to ensure that less listing of defaults expected in experimental groups would not simply be an artifact of their having fewer total values than defaults. (Recall, default attributes took on only two values within a set, one for each category, whereas variable attributes took on four different values, with two different values for each category). Unlike the previous two experiments, which used Stanford undergraduates as subjects, the subjects in Experiment 2A were recruits from Lackland Air Force Base.

As shown in Fig. 6, evidence of subcategory learning was obtained in the experimental groups. Subjects tended to reduce their listing of defaults and increase their listing of variables as more stimuli were encountered. When listing of defaults is subtracted from listing of variable attributes, it can be seen that the learning effect in these conditions is quite substantial; the difference in responding increased from -2% on the first trial to 48% in the last trial (see Fig. 6, third panel). The increasing trend in this index is statistically significant \( r(19) = 2.84, p < .02 \). By contrast, no such trend appeared for subjects exposed to the fully variable, "control" stimuli. When listing of defaults (two-valued attributes) was subtracted from that of variables (four-valued attributes) for control subjects, no significant learning effect was observed \( r(23) = 0.92, p > .20 \). Moreover, in direct comparisons, defaults were listed an average of 25% less often by the experimental subjects than were the corresponding features by the control subjects \( r(46) = 22.39, p < .001 \). Thus, despite the intermixed sequence of training instances, significant learning, reflecting correlations in the default values defining the two categories, occurred in this experiment.

These results may be compared to those of Experiment 2 to examine how the different sequencing of training instances affected learning. Comparing the learning curves obtained in these two experiments (Fig. 6 vs. Fig. 5), learning occurred faster and more clearly in Experiment 2 than in 2A. In addition, the asymptotic listing of defaults is far greater in Experiment 2A than in Experiment 2. Comparing the listing of defaults in the final, mixed, block in Experiment 2 (trials 32-40) to default listing in the same trials from Experiment 2A, listing was 16% higher in the mixed sequence compared to the blocked sequence. Comparison of the two experiments, matched by trials, yielded a significant difference \( r(7) = 9.63, p < .001 \).

Unfortunately, this difference cannot be interpreted as unambiguous evidence for interference in the mixed condition. Technically, it is inappropriate to compare data across different experiments conducted at different times with different groups of subjects. This problem arises in this comparison because the subjects in Experiment 2 were Stanford undergraduates whereas those in Experiment 2A were Air Force recruits. The different levels of learning could have been due to different average
learning abilities in these two subject populations, as well as the training sequences. Obviously, a more appropriate procedure would be to compare blocked and intermixed conditions within one population and experiment. (We should note, however, that several pilot versions of this mixed-sequence experiment were presented to Stanford undergraduates as we were developing the version administered to Air Force recruits, and learning by these pilot subjects always appeared much poorer than was observed in Experiment 2). In light of the absence of a direct comparison between different sequencing conditions to substantiate the interference hypothesis, the conservative conclusions from this experiment are that subjects clearly can learn contrasting categories even from mixed training sequences and that attribute listing provides a useful index of this learning.

D. EXPERIMENT 3: UNSUPERVISED LEARNING OF A HIERARCHY OF CATEGORIES

In accumulating knowledge about a domain, people often develop a set of related categories at multiple levels of specificity. Many real-world domains, such as categories of animals, plants, automobiles, jet aircraft, or medical diseases, are partitioned at more than one level as some form of default hierarchy. One way domain experts differ from nonexperts is by the rich conceptual hierarchies of interrelated subcategories they have acquired, as well as their facility in using this knowledge to improve their processing and retrieval of new information in the domain (e.g., Holland, Holyoak, Nisbett, & Thagard, 1986). Given the prevalence and importance of such conceptual hierarchies, it is odd that prior research on category learning has usually examined single-level categories. There have been few demonstrations of learning of multilevel categories or even reliable methods for observing such learning (but see Murphy & Smith, 1982).

Experiment 3 was intended to demonstrate that subjects could spontaneously induce categories in a multilevel domain and that the attribute listing procedure could track this learning. In contrast to previous categorization models suited only for learning single-level partitions, our theory can apply to the progressive learning of categories and subcategories at multiple levels of specificity. Multilevel learning is possible because the model assumes that once defaults at a given level are learned, the subject will take them as “background” and proceed to focus on other aspects of learning.

Fig. 6. Percent listing of defaults and variables from the experimental group in Experiment 2A. The third panel shows the difference in listing between defaults and variables on each trial for the experimental group, while the bottom panel provides the corresponding index for the control group.
the stimulus. Such refocusing is conducive to finding previously unnoticed, correlated attributes of the stimuli, leading the model to attend to features that might form a basis for learning more differentiated subcategories.

Experiment 3 was similar to Experiment 2 except that instead of two contrasting categories, four categories of exemplars were presented in blocked sequence. Instances of the first insect category — call it A1 — were presented for the first 10 trials, followed by 10 A2 instances, then 10 B1s, and 10 B2s. Each insect varied in eight attributes. The default values characterizing the four categories can be denoted as follows: 

\[ A1 = 111111XX, A2 = 111222YY, B1 = 222333QQ, \text{and } B2 = 222444RR, \]

where X, Y, Q, and R denote different pairs of values of variable attributes occurring in each of the four categories. The superordinate defaults (A vs. B) occur on the first three attributes, whereas the subordinate defaults are defined by the values of the fourth, fifth, and sixth attributes. (The blocked sequence was intended to facilitate the learning of the category discriminations in this initial demonstration experiment; later studies can examine the boundary conditions of such learning.) For testing purposes, following the four training blocks, a mixed block was finally presented in which instances of all four categories were presented in random sequence. (A control condition was also included; subjects in this group received stimuli constructed from random combinations of the two-, four-, and eight-valued attributes.) The listing task was similar to those of previous experiments, except that subjects were explicitly told to limit the number of features listed by imagining that each listed feature would cost them 25 cents, whereas each incorrect identification (based on their lists) on a final recognition test would cost them one dollar.

Turning to the results, the pattern of responding generally conformed to the model's predictions (see Fig. 7). For superordinate defaults (top panel of Fig. 7), listing decreased from 49 to 9% as successive A1 instances were encountered, and decreased further during the A2 block. This decreasing trend occurred for 10 of 11 subjects, and was significant at the .01 level by a sign test. Presenting the first A2 instance caused an abrupt drop in listing from a low level of around 9% to near zero. This drop may reflect attentional competition at this point, since the surprising features of the first A2 instance may have shaken subjects out of a routine pattern of listing some A defaults unnecessarily. A 15% increase in listing of superordinate de-
faults was then observed on the first B1 instance (marginally significant, $t(10) = 2.18, p < .10$), followed by a rapid decline back to the zero baseline [$t(10) = 2.23, p < .05$]. Listings did not increase significantly on the first trial of the mixed block [$t(10) = 1.0, p > .20$]; this indicates that subjects had learned stable discriminations at the superordinate level during the earlier training blocks.

Listings for subordinate defaults were also consistent with the model (see middle panel of Fig. 7). Listings declined significantly from 52 to 95% over trials for the A1 category [$t(10) = 4.31, p < .01$] and then increased sharply on the first instance of each of the following categories. These increases were statistically significant for A2 [from 9 to 85%, $t(10) = 11.69, p < .001$] and B2 [6 to 30%, $t(10) = 2.66, p < .05$], but not for B1 [15 to 27% $t(10) = 1.31, p > .20$]. The lack of significance for B1 may have been due to the fact that defaults at both levels (attributes 1–6) shifted on this trial, and this competition may have reduced the number of defaults at each level that might otherwise have been reported. As expected, on the first trial of the mixed block, reporting of default values did not increase [$t(10) = 1.49, p > .10$]. Moreover, listing of these values by this time was 80% lower than listing of variable attributes [$t(10) = 17.53, p < .001$]. Throughout the mixed block, listing of defaults was far lower than that of comparable four-valued attributes in the random control group [by 62%, $t(7) = 9.59, p < .001$].

Thus, subjects seem to have learned stable discriminations at both the superordinate and subordinate levels. Moreover, subjects apparently transferred superordinate defaults learned for A1 to A2, and for B1 to B2, since no increase in listing of superordinate values was observed when instances of the latter subordinate categories were first introduced. The only unexpected result was the higher listing of variable attributes than of default violations on the first trial of the later blocks (B1 and B2). Possibly subjects developed a fixed routine of listing selected variable attributes during these early blocks and simply continued this routine into later blocks. Since the variable attributes were uniquely identifying in this experiment (i.e., variable values predicted categorization since a different pair of values were used for each category), such selection would be a reasonable listing strategy.

Setting aside this small discrepancy, the results clearly show that subjects could learn without supervision to distinguish multiple categories in a hierarchically organized domain. More research will be required to characterize details of how this knowledge is organized in subjects’ memories, and to identify the boundary conditions and major variables that influence such learning.

E. EXPERIMENT 4: SIMILARITY RELATED TO ATTRIBUTE SALIENCE

A basic assumption of our approach is that attentional salience will be controlled by informativeness. Subjects’ patterns of attribute listings in the preceding experiments could be interpreted as an index of attentional salience, but subjects in those experiments were directly instructed to report the most discriminating features of the instances. While subjects’ listings provide a good record of changes in their beliefs about a category, it is not obvious that they provide an index of salience that would generalize to other situations. A more valid measure of attentional salience per se might be provided by people’s judgments of the similarity of two patterns from the same general class. In applying our model to predict similarity ratings, we will assume that presented stimulus pairs are first encoded into working memory, and that these working memory representations constitute the input to the comparison process. Because informative features are expected to have high salience in such memory representations, they should exert a strong effect on ratings of similarity. In general, subjects’ judgments of how similar two patterns are should be dominated by the instances’ variable or exceptional features, whereas category defaults should have little impact on subjects’ comparisons.

The hypothesis that learning category defaults reduces their impact on similarity judgments leads to the seemingly paradoxical implication that two instances of a category may become more dissimilar as people become increasingly familiar with that category. A moment’s thought reveals that this is actually a commonplace phenomenon associated with expertise: people who become very expert about a particular domain (e.g., expert botanists, wine tasters, dog-show judges, etc.) become highly sensitive to differences among the objects in that domain while taking for granted their well-known commonalities. For instance, a black oak and a red oak are much less alike to an expert botanist than they are to most nonexperts. In fact, people’s tendency to take for granted what they know about a general domain and to focus attention on the novel aspects of instances is probably what allows experts to discover progressively more specific subcategories within that domain. It may also be a powerful factor that promotes perceptual learning, i.e., improvements in perceptual judgments and discriminations that accompany experience in a given domain (see e.g., Gibson, 1963, 1969).

Another implication of this analysis is that when an instance violates a default expectation of its category, that violation should be highly salient and have a strong impact on similarity. We can use this hypothesis to
predict circumstances in which decreasing the number of common features shared by two stimuli leads paradoxically to increased similarity—an implication in direct contradiction to results ordinarily obtained in similarity experiments (e.g., Gati & Tversky, 1984). If presence and absence are conceptualized as two alternative values of a binary attribute, then as one value occurs more frequently within a category, its informativeness is decreased whereas the informativeness of the other value is increased. Thus, the absence of a highly expected default from a given instance should appear more informative than its presence. If this surprising absence occurred in two instances being judged for similarity, it would have the paradoxical effect of increasing the salience of their “common” features. By similar reasoning, deleting a default feature from one instance but not from another would result in an unusually salient difference between them, greatly reducing the similarity between the exceptional and unexceptional instances.

The experiment to test these implications consisted of a series of similarity judgments in which college student subjects rated the similarity of pairs of instances on a 20-point scale. The stimuli were realistic line drawings of fictitious insects (“bugs”), similar to those used in the attribute listing experiments described above. Several features (e.g., wings, tails, antennae) could be added or removed to construct different instances. Two of these features were consistently presented in all instances (defaults), and the others were presented half the time (variables); in addition, instances were varied along several other attributes to increase the perceived variability of the category. We expected that after having seen several pairs of bugs (with no category feedback whatever), subjects would learn structural norms for the consistently correlated features, treating them as a category of stimuli. These norms would specify which features were correlated (expected defaults) and which tended to vary across instances. In the midst of this uniform training series, stimulus pairs were occasionally presented in which one or both insects violated the category expectations; such bugs would either be missing an expected default or possess an extra feature not seen in any of the other instances. We were interested in how subjects would rate the similarity of two bugs that were deviant in the same way, in contrast to the way subjects rated matched, “normal” pairs of bugs.

As expected, the results showed that subjects’ expectations influenced their similarity judgments. However, violations of defaults had a much larger effect when they served as distinctive features (differences) than when they served as common features. As predicted, pairs in which one member was missing an expected feature (or in which a previously encountered feature was added) were rated significantly less similar than pairs that differed by a variable feature. To illustrate, if one insect had wings and the other did not, the effect of this difference on perceived similarity was greater if subjects expected all instances to have wings (2.60 points) than if they expected wings to be present or absent equally often (1.42 points); the difference between these two effects was statistically significant \( t(21) = 3.20, p < .01 \). However, contrary to predictions, when both test instances were missing an expected feature, their similarity was not increased by this shared anomaly; such pairs with missing defaults were rated as equally similar as pairs in which the defaults were present \( t(21) = 0.48, p > .50 \). To illustrate, if wings were an expected default, then two bugs that had wings were rated as similar as two bugs that did not. But for subjects who learned wings as a variable feature, pairs that shared this attribute appeared slightly more similar (by 0.22 point) than pairs in which it was absent \( t(21) = 2.84, p < .01 \).

Although these results clearly showed that subjects’ category norms influenced their judgments, we were disappointed that pairs lacking expected defaults were not rated as more similar than normal pairs. Perhaps this was due to subjects’ weighing distinctive features more than common features in their pairwise similarity judgments [over six times as much, \( t(21) = 10.80, p < .001 \]. This greater weighting of distinctive features is the typical result with pictorial stimuli (see Gati & Tversky, 1984). Indeed, subjects’ reports (and other data) indicated that most of our subjects were computing similarity of two bugs by simply counting the features that differed between the bugs, and largely ignoring their common features. Such a difference-counting strategy would, of course, wash out the impact of our manipulation of common features. To circumvent this strategy, we designed a second study in which we could pursue the “common deviation” effect in a situation that minimized those strategic factors that mitigated its appearance in Experiment 4.

F. EXPERIMENT 5: SIMILARITY OF MEMORIZED INSTANCES

In this experiment, subjects were asked to rate from memory the similarity of instances, given only verbal labels designating the bugs they had learned earlier. We expected this modified procedure to have several advantages over similarity ratings of explicitly displayed instances. First, due to memory limitations it should be more difficult for subjects to use an artificial, attribute-by-attribute, differencing strategy, as they apparently did in Experiment 4. Instead, subjects should be more likely to make their ratings intuitively, from a holistic impression of the instances’ similarity. Second, because people’s memories of instances tends to be dominated by their unusual, unexpected features, similarity judgments from memory
may be more influenced by such exceptional features than would judgments of displayed stimuli. Third, comparisons from memory are arguably more natural and interesting in some respects than comparisons between displayed instances. Similarity in memory is probably an important factor controlling categorization, spontaneous reminders across separate episodes (see Ross, 1984; Schank, 1982), and the formation of novel subcategories based on informative commonalities between specific instances (Malt, 1989).

This experiment consisted of two phases. In Phase 1, subjects learned to associate a specific label (a CVC nonsense syllable) with each of 10 instances from a single category of insects. The insects were constructed from the same materials used in Experiment 4. For each subject, eight of the presented instances possessed a target default value, whereas this value was absent from the remaining two instances. To balance stimulus-specific effects, four different stimulus sets were constructed and presented to different groups of subjects. Each set had a different default feature that was absent from 2 of the 10 instances. Thus, the influence on similarity judgments of presence vs. absence of a default feature was compared for four defaults across the experiment as a whole. Subjects were taught the names of the 10 instances by a cued recall procedure. They were first shown a given stimulus, were asked to label it, and were then told its correct label. This training continued until subjects could correctly name all the instances, or until they had completed 20 cycles through the 10 instances.

In Phase 2, subjects rated the similarity of specific pairs of insects learned earlier, referred to only by their CVC labels. Ratings were made on a 20-point scale, where 1 indicated very low similarity and 20 indicated very high similarity. Two of these pairs referred to the two deviant instances that were lacking a given target default. For each such target pair, another pair was presented that was identical to the target pair except that the target default was present in both instances. Several filler pairs were also included, which varied in their number of mismatching attributes. If subjects were attending to the task and rating similarity in a manner consistent with previous research, pairs would be rated as more similar the fewer the differences between the two insects.

As predicted, the results showed that increasing the number of mismatching attributes in a pair reduced similarity, $t(16) = 10.42, p < .01$. Each mismatch between the two members of a pair decreased their rated similarity by an average of 3.52 points on the 20-point scale. Thus, the effects of mismatching attributes demonstrated that the rating-from-memory procedure produced an overall pattern of results comparable to those found in standard similarity experiments.

The more interesting data concern the similarity of pairs for which the default attribute is absent. As predicted, removing a default from both members of a pair increased their rated similarity by an average of 3.35 points above the rating given to control pairs in which that default value was present—a statistically significant effect, $t(16) = 2.67, p < .02$. Thus, the predicted effect of surprising attribute values (absence in this case) on similarity was confirmed by these data.

One difficulty with the similarity-from-memory procedure, however, is that it is not clear to what extent outcomes reflect the way in which instances are encoded during the training phase, in contrast to the way in which they are later retrieved and compared during the similarity rating phase. For example, our model predicts that subjects would learn each instance by recording its category membership and then learning mainly features that are highly informative with respect to that category (i.e., variables and missing defaults). However, if subjects focused on absent defaults in learning the unusual instances, they might have allocated correspondingly less attention to the variable features of these instances (relative to the amount of attention these features would receive in normal instances). In fact, subjects need not have learned all the variable features of the unusual instances, since the necessary discriminations could have been acquired by learning only the absent default plus a single variable feature to distinguish between the two unusual instances. Thus, it is not clear whether the greater similarity of pairs sharing an absent default was due to the greater salience of this unusual value at the time of comparison or to fewer differences between the pair members available from memory.

One way to eliminate this ambiguity would be to use a training procedure that forced subjects to learn all the features of each instance well enough to ensure that they would be available from memory during the similarity ratings. This could be accomplished by using a training procedure in which subjects recalled all the features of instances when cued with their names, rather than the reverse, recalling the name to the presented bug, as in the present experiment. Such a modification would force subjects to learn all the features of each instance. If later similarity ratings still showed an effect of shared absence, it could not be explained by fewer differences between the unusual pairs available from memory. If the shared-absence effect was eliminated by such a learning regimen, then shared absence would be considered an encoding effect rather than reflecting comparison strategies per se. But because such encoding biases are assumed to influence much of people's learning of real-world stimuli, they could indirectly affect many phenomena related to similarity, such as reminders, the discovery of new subcategories, and generalization of learning. Our model is quite compatible with an encoding explanation as well as a salience-in-comparison explanation of the shared-absence effect.
G. EXPERIMENT 6: INSTANCES STORED IN RELATION TO CATEGORIES

The aim of the next experiment was to study the impact of a category schema on learning and memory performance in processing specific instances. According to our theory, people should be biased to record into memory new instances of a well-known category by learning primarily their informative features, while bypassing their uninformative defaults. By storing the two types of features at separate locations in memory (defaults associated with the schema, nondefaults with individual instances), the two feature bundles would not interfere with each other's retrieval. Presumably, such learning benefits are a major reason that experts in a given domain can process and remember stimuli in that domain much more efficiently than nonexperts (e.g., Chase & Simon, 1973; de Groot, 1965, 1966).

The experiment conducted to demonstrate these advantages due to schema-based encoding consisted of three phases. In Phase 1, subjects were pretrained on the features characteristic of several categories. The categories were types of astronomical stars, supposedly differentiated by their chemical compositions. That is, a category (e.g., blue stars, red stars) was coordinated with a list of chemical elements characteristically found in that type of star. Attempting to simulate the easy command people have of their knowledge of everyday categories, we trained subjects on these feature lists until they could remember them very easily. In Phase 2, the subjects learned several named instances of each category (stars such as Rigel and Vela). These named instances were described solely in terms of their chemical constituents, which contained all the features true of the general category to which they belonged, plus one or more variable features distinctive to that instance. The number of default features in a given category (two to four) was orthogonally varied along with the number of extra variable features attributed to particular instances (one or three). In Phase 3, subjects were tested on their knowledge of (1) the features associated with particular named instances and (2) the categories to which such instances belonged. The testing was a speeded verification task, in which instance questions (e.g., "Rigel contains hydrogen") had to be judged true or false as rapidly as possible while maintaining accuracy. Reaction times were the major dependent measure of interest.

Our main predictions were based on the familiar "fan effect" (see Anderson, 1976), which implies that the more features associated with the tested instance, the longer subjects should take to verify any particular one. The predictions can be easily understood by referring to Fig. 8, which illustrates two possible memory organizations in terms of an associative network notation. In Fig. 8B, the categorization of the instances is directly

---

Fig. 8. Two possible memory organizations relating general information about categories to specific information about instances.
encoded, but no direct associations have been formed between the instances and the category defaults. However, those defaults are still accessible from the instances via indirect retrieval (interference) through the category. The key predictions of this model concern the effects of the number (fan) of default vs. variable features on verifying variable-feature probes. Because of the partitioning of subnodes in Fig. 8B, the number of defaults should exert no fan effects on the retrievability of instance-to-variable feature associations. By contrast, the time to verify a variable feature should increase with the fan of variable features associated to the instance.

Figure 8A shows an alternative hypothesis that assumes that subjects store common features of the category redundantly and directly with each of its instances. This is a necessary assumption of pure exemplar-storage models, which assume that people represent category knowledge merely by storing instances and then compute generalizations about the category from those instances as needed. In such a representation, default properties of a category could only be inferred if they were stored in a large proportion of its memorized instances. According to this hypothesis, category defaults will be directly associated with the instance, so they should produce fan effects on verifying variable features. That is, increasing default fan should cause the same slowdown in verifying variable features as would increasing variable fan, in contrast to the segregated organization illustrated in Fig. 8B.

Turning to the reaction time results, a robust fan effect on verifying variable features was demonstrated as a function of the fan of variable features associated with the tested instance. That is, the time to verify an instance-to-variable feature question increased in the predicted manner with the instance’s number of variable features \( F(1, 14) = 28.56, \ p < .0001 \). However, the number of default features inherited from the category had no significant effect on the speed with which variable-feature probes were verified \( F(2, 28) = 0.84, \ p > .25 \). It is remarkable that increasing the number of category defaults produced an insignificant slowdown in retrieving variables because these defaults were verified at least as fast as the variable features, and hence (according to the theory in Fig. 8A) should have produced equally strong interference if they had been directly associated with the instances. In a direct comparison, the 588-msec variable fan effect on verifying variable features was reliably greater than the nonsignificant 215-msec default fan effect, \( t(14) = 2.27, \ p < .05 \) (reaction times to confirm variable features averaged 2634 msec in this experiment). Moreover, greater default fan did not significantly slow down recognition of instance-to-default probes \( F(2, 28) = 0.98, \ p > .25 \), whereas increasing variable fan did slow down recognition of the instance-to-default probes \( F(1, 14) = 5.94, \ p < .05 \), further confirming the dissociation between default and variable fan effects. Also, subjects were over 800 msec faster to verify the categorization of instances than either instance-to-category feature or instance-to-variable feature probes \( F(2, 28) = 36.70, \ p < .0001 \), even though they had never been explicitly taught this information. This result suggests that subjects noticed, rehearsed, and made strong use of category membership in encoding the original instance information, as predicted.

The results of this experiment are consistent with the assumption that people can learn summary models of categories, and that these models play a strong role in determining the encoding of instances and their organization in memory. The results are incompatible with extreme versions of exemplar-storage theories, which assume that both default and variable features are stored together in association with specific instances (see Fig. 8A). In addition, the results validate the utility of the fan effect technique for investigating issues related to the abstraction and organization of category knowledge.

H. EXPERIMENT 7: DEFAULTS IMPROVE LEARNING OF VARIABLES

Experiment 6 provided evidence that people encoded instances in terms of their category membership, which would have produced the memory structure depicted in Fig. 8B, rather than recording a full listing of their features, which would have created the structure depicted in Fig. 8A. But the subjects in that experiment were directly taught the generalizations they had to know for later use. It is important to check whether similar encoding processes occur when the general schemas are being learned concurrently with the instances. Therefore, a second memory task was developed to observe how people learn and apply category models for themselves, using only instances without prior explicit category training. In this experiment, subjects induced for themselves the shared properties that defined categories of stimuli. We then examined the organization of the resulting memory and how the categories were used for encoding new instances.

The procedure involved presenting a training instance on each trial on a CRT screen. An instance consisted of a series of letters; for example, a particular instance might contain the letters B, D, Q, and N. This feature list remained on the computer screen for a brief period (1 sec per feature) during which the subject was asked to study and try to memorize it. After this brief study period, the letter string disappeared from the screen and was followed by a short distractor task to reduce short-term memory. This
The training instances (letter sequences) fell into three distinct categories, two of which were characterized by correlated elements and the third of which was a "junk category" consisting of randomly generated sequences of different lengths. For the two categories with correlated features, there was a specific set of letters that consistently appeared in the same positions in every instance, with different consistent letters for each category. As in Experiment 6, the fans of default and variable elements were independently manipulated across different instances. Instances of one category had three consistent default letters while instances of the other category had four defaults. In addition to the default letters, each instance contained either one or two extra, variable letters. For purposes of comparison, an equal number of randomly constructed control stimuli were presented as a junk category. These control stimuli were matched with those from the correlational categories in their number of letters but had neither of the default letter clusters. Subjects were not told which category each instance belonged to, or even that there were separate categories in the experiment.

The training instances and categories in this experiment were designed in a similar manner to those in the previous experiments, except that the stimuli were letter strings rather than pictures of objects (insects) or lists of verbal descriptors. Letter string stimuli allow subjects to conveniently record their instance memories on the computer keyboard. From the perspective of our model, the use of letter strings as stimuli should make little difference because they are as easily described in terms of attributes (serial positions) and values (the specific letters appearing at each position) as other types of category materials. Moreover, categories are characterized in exactly the same way as in previous studies, namely, in terms of intercorrelated, mutually predictable features. Furthermore, letter sequences have often been used as materials in standard studies of classification learning (e.g., Bower & Trabasso, 1964; Hunt, 1962; Rosch & Mervis, 1975), and generally behave much the same as other types of stimuli used in artificial category experiments.

We expected that after repeated experience with the training instances, subjects would learn which groups of features consistently co-occurred across instances. Once an instance was recognized as containing a familiar cluster of correlated features, i.e., as belonging to the category character-
with those from Experiment 6, which showed that variable features were encoded mainly as distinctive properties of specific instances whereas defaults were stored with the generic schema for a given category. They provide further evidence against pure exemplar-storage models which assume that subjects must encode and store both the default and variable properties of category members. Furthermore, the success of the experimental task provides opportunities for detailed investigation of the structure of these category models, how they are learned, and the strategies by which they are applied.

I. EXPERIMENT 8: VARYING RELIABILITY OF DEFAULTS

In most realistic learning situations, the data on which learners must base their generalizations contain errors and exceptions. Many beliefs about natural categories are violated by specific instances. For example, the ability to fly is a default property of birds in general, but several birds (e.g., ostriches, kiwis, penguins) violate this default. The notion that natural categories can be defined in terms of necessary and sufficient features has come under vigorous attack (see Rosch, 1975, 1977; Smith & Medin, 1981), and the view that categories are “fuzzy” with probabilistic features has been promoted. Therefore, it is important to ask whether the outcomes of our previous experiments would hold for situations in which category defaults occurred with moderate to high degrees of unreliability. The purpose of the following experiment was to study people’s ability to learn and apply schemas based on “noisy” input data, for which generalizations would be somewhat unreliable in that usually consistent features would sometimes be replaced or missing from particular instances.

A recall procedure similar to that in Experiment 7 was used. The stimuli were sequences of six letters; each position in the sequence can be thought of as an attribute, and the letters filling that position serve as the alternative values of the attribute. Three possible letters could occur at a given position, providing 3⁶ possible stimulus patterns. On each trial, subjects were presented with a single instance (six-letter string) for a brief study period, followed by a 15-sec distractor task to reduce short-term memory (adding or subtracting digits from a running total). They then tried to recall in any order the six letters presented on that trial. Each subject was presented with instances of a single category, characterized by both consistent and variable attributes. For the consistent attributes, one value occurred more frequently than the other two; this was the modal (default) value of that feature. In contrast, all three values of the variable attributes occurred equally often. The major independent variables were (1) the probability of the modal value of the default attributes (60, 70, 80, or 90%) and (2) the ratio of default to variable attributes characterizing the concept (four defaults and two variables vs. two defaults and four variables). A control condition was also included in which all six attributes were completely variable. Thus, a total of nine conditions were tested in a between-subjects experiment design. The subjects were 227 Air Force recruits from Lackland Air Force Base, who were randomly assigned to nine experimental groups.

We were interested in how default reliability, and the ratio of default to variable attributes, would affect people’s ability to learn and apply default schemas. We predicted that subjects would be able to learn such schemas even when they were fairly unreliable (i.e., at lower levels of default probability and default/variable ratios). However, the degree to which subjects could use the schema to improve their learning of its instances should depend on several factors. First, the proportion of attributes with strong defaults should determine how much attentional capacity can be allocated to encode the remaining, nondefault values. Second, as the probability of a category’s defaults is decreased, their strength in the category norms also declines. As a result, their perceived informativeness will increase, attracting more attention. This increases competition for attention among the other features, reducing the benefits of schema abstraction for recall. Third, the poorer the fit of an instance to its schema, i.e., the more exceptions it contains, the less the schema will facilitate learning of that instance. Each exception draws more attention than the default it replaced, increasing its share in the competition for attentional resources. To summarize these considerations, the schema theory predicts that performance should be highest for subjects whose instances display the highest level of predictability (i.e., instances with no exceptions, a high ratio of defaults to variables, and default values occurring with 90% probability), and performance should decrease as predictability is decreased. The poorest performance should occur, of course, for control subjects who see only randomly generated letter strings.

The results of Experiment 8 strongly supported these implications. Taking the highest predictability group as a reference standard, all three independent variables significantly affected recall. Increasing the number of exceptions per instance significantly decreased the recall of both variable and default features [F(2, 424) = 12.35, p < .01, F(1, 212) = 25.39, p < .01, respectively; see Fig. 9, left panel]. The higher the default probability, the higher the recall of variable features [F(4, 60) = 8.65, p < .01] and defaults [F(4, 60) = 11.56, p < .01; see Fig. 9, right panel]. The ratio of the number of defaults to variable features also affected average recall. When default values were 90% reliable, variable features were recalled 10% better in the high-ratio condition than in the corresponding low-ratio
condition \( t(24) = 2.42, p < .05 \); defaults were also recalled about 10% better in the high-ratio group \( t(24) = 1.97, p < .10 \). As predicted, the data indicate that subjects learned default expectations and used them to facilitate recall even for noisy input data. For example, although recall of defaults in the high-ratio groups declined with decreasing default probability, even the 60% group showed evidence of improved recall due to schema learning relative to the control group that saw only random letter sequences [a 10% difference, \( F(21) = 1.91, p < .10 \)].

These results confirm several predictions of the schema model and show its applicability to categories with probabilistic features. The fact that variable features as well as defaults were better recalled at high levels of predictability suggests that category knowledge had a top-down influence on encoding process beyond simply helping subjects to guess default values on the recall test (see discussion in Experiment 7). The results are incompatible with strictly bottom-up learning theories, such as pure exemplar storage models.

VI. Concluding Comments

Overall, the studies yielded results highly consistent with the proposed theoretical framework. The attribute-listing experiments revealed patterns of unsupervised learning characterized by initial discovery of categories based on featural contrasts with previous default expectations, gradual learning of defaults within a new category, and an overall bias to report informative (exceptional and variable) features of the instances. The task also reveals the distinct stages in learning categories in hierarchical domains. Norms about a prior source category transferred to new derivative categories when they were applicable. The attribute-listing task shows promise as a useful paradigm within which unsupervised learning can be studied in complex, hierarchically organized domains, and in which detailed theories may be tested and refined.

The similarity experiments provided independent confirmation of the proposed attentional biases toward surprising features. In particular, the similarity-from-memory procedure appears promising for investigating factors determining perceived informativeness. Such tasks may provide greater insights into comparative judgments themselves, by suggesting constraints regarding the representation of stimuli in such tasks.

The memory studies provide further support for the model's attentional assumptions, and explicate their impact on encoding processes, organization of category and instance information in memory, and later retrieval of facts about instances. The results of those experiments disconfirm an assumption of simple exemplar storage theories, i.e., that both predictable defaults and informative nondefaults are encoded together as properties of individual instances. In concert with our earlier tasks, the memory paradigms may prove useful for investigating fundamental issues related to domain expertise and its cognitive benefits. The results of such experiments also suggest how people's prior stereotypes influence their processing of and memory for events and objects they encounter.

We have attempted to sketch a general model of learning that describes the abstraction of observations into general concepts and then the utilization of these concepts to encode later instances. Since top-down (conceptually driven) and bottom-up (data-driven) components are closely intertwined in human learning, this approach has some advantages compared to strictly bottom-up or top-down approaches. Equally important as our specific theoretical formulation are the general methods or paradigms developed to investigate an important set of new issues. For example, unsupervised learning, especially in multilevel domains, has been little investigated in previous research; also, little prior research has analytically examined schema-based effects on new instance learning by synthesizing and manipulating schematic knowledge in the laboratory.

Many extensions of the methods discussed may be suggested. For example, variants of the attribute-listing task could provide further converging observations of unsupervised learning and shifts in attentional biases under various task constraints. Variants might include (1) asking subjects to rate directly the surprisingness or distinctiveness of each attribute on
each trial; (2) recording how much time a person devotes to looking at each instance and correlating this with the number of informative values the instance contains; (3) using text descriptions of the instances, recording line-by-line reading times, and relating those to the predicted informativeness of each statement with respect to a general schema; or (4) requiring two subjects to communicate with one another about successive instances (Krauss & Weinheimer, 1966) and noting how their referential descriptions of stimulai change as they acquire shared knowledge of the stimulus defaults. But confining ourselves to the present set of tasks, we have only begun to investigate many relevant variables, such as interference among related categories and variations in feature distributions, that could provide further tests of our model of unsupervised learning. Such issues should provide fruitful topics for future investigation.

ACKNOWLEDGMENTS

This research was supported by the Air Force Office of Scientific Research Grant #87-0282. We express our appreciation to the personnel at the Air Force’s Learning Abilities Measurement Program (LAMP) at Brooks Air Force Base in San Antonio, Texas, for the use of their facilities and subject pool in running several of these experiments.

REFERENCES


Category Knowledge in Unsupervised Domains
