Category Invention in Unsupervised Learning

John P. Clapper and Gordon H. Bower

This research aimed to discriminate between 2 general approaches to unsupervised category learning, one based on learning explicit correlational rules or associations within a stimulus domain (autocorrelation) and the other based on inventing separate categories to capture the correlational structure of the domain (category invention). An "attribute-listing" paradigm was used to index unsupervised learning in 3 experiments. Each experiment manipulated the order in which instances from 2 different categories were presented and evaluated the effects of this manipulation in terms of the 2 competing theoretical approaches to unsupervised learning. Strong evidence was found for the use by Ss of a discrete category invention process to learn the categories in these experiments. These results also suggest that attribute listing may be a valuable method for future investigations of unsupervised category learning.

The study of concepts and category learning has long been a focus of research in cognitive psychology. Most of this research has studied supervised category learning, in which a tutor provides the subjects with category labels and feedback relevant to the success criterion of the learning task (e.g., Bruner, Goodnow, & Austin, 1956; see Millward, 1971, for a review). By contrast, unsupervised learning has received much less attention by experimental psychologists. In unsupervised learning, subjects must invent and use categories without predefined category labels or feedback from an external tutor. Many categories that people learn in real life are acquired in observational, untutored conditions and thus are examples of unsupervised learning. Much of our knowledge about the properties and behavior of common physical objects, social interactions, linguistic classes and rules, and everyday tasks and procedures may be learned in this manner (Billman & Heit, 1988). Any learning by pioneers about a novel environment is unsupervised because they must invent their own categories for describing that environment and generate their own criteria for classifying stimuli into these categories.

This article describes a recently developed procedure for investigating unsupervised learning (see Clapper & Bower, 1991) and three experiments in which this procedure was used to test theories of how categories could be learned and represented in unsupervised tasks. We begin by describing more precisely what we mean by an unsupervised learning task and how categories could be defined within such a task. We then argue that models of unsupervised learning can be divided into two general types, which differ in how category knowledge is represented in long-term memory and the processes by which this knowledge is abstracted from individual training instances. After providing this background, we describe the attribute-listing paradigm and show how it can be used to discriminate between the two classes of theories described earlier.

Defining Categories in Unsupervised Tasks

In supervised learning tasks, categories are predefined by the experimenter and subjects must use the experimenter's feedback to determine the correct rules for assigning stimuli to each category. Any arbitrary categorization rule may be used in such experiments (e.g., disjunctive rules such as "Members of Category A are either red squares or blue diamonds, but not red diamonds or blue squares"), and categories need not be functionally natural or capture informative patterns within the stimulus set. In contrast, in unsupervised tasks categories are not arbitrarily predefined by an external tutor; rather, subjects must discover categories for themselves as they explore a given stimulus domain. This presumably requires that some regularity or structure actually exist within that domain, that is, a pattern or signal that can be distinguished from the noise of background stimulus variation. It is necessary to define what kind of pattern or structure constitutes a category before proceeding to evaluate whether subjects in a given condition have learned this category.

Following Clapper and Bower (1991), we adopt a conventional feature-based vocabulary for describing commonalities and differences within a stimulus set and then define categories in terms of this vocabulary. Individual stimuli are described as collections of features. Each feature can be thought of as a specific, concrete value of a more generic or abstract attribute. For example, the stimuli in a given set could be described in terms of their shape (a generic attribute), with particular stimuli being squares, circles, or triangles (the specific values of the shape attribute). In principle, the values of an attribute could be either discrete (e.g., squares vs. circles) or continuous (i.e., ordered quantities, such as gradations of size or shading), but only the discrete-valued case is considered here.

Given a set of attributes for describing a stimulus domain, patterns of correlated features (attribute values) provide an inductive basis for partitioning that domain into subsets or
categories. To illustrate, a collection of fruit flies bred in a geneticist's laboratory could be described in terms of attributes such as size, eye color, wing shape, leg length, and so on. If it was then observed that individuals with long wings also had red eyes, large size, and long legs, whereas those with short wings had white eyes, small size, and short legs, these patterns of feature co-occurrences would form an inductive basis for recognizing two distinct categories of fruit flies within that population (Clapper & Bower, 1991).

Figure 1 shows several stimulus sets with different types of correlational patterns that could serve as a basis for partitioning them into separate categories. Within each of these sets, some attributes have strongly correlated values whereas others do not. For example, in Stimulus Set 1, the first five attributes listed have perfectly correlated values whereas the last three attributes vary independently. We refer to correlated values as default values of the category to which they give rise. Attributes that are uncorrelated within a given category are referred to as variable attributes.

1 Of course, the existence of such patterns depends on the particular set of attributes used to describe a given stimulus. Thus, the same set of stimuli might be categorized differently with respect to different sets of attributes. In principle, the categorization of a given stimulus and the attributes used to describe it are somewhat mutable and dependent on the task context and the other stimuli with which it is contrasted. In practice, experimenters usually define a set of canonical attributes by which a stimulus set is generated and described, and this determines the normatively "correct" categorization of that set to which subjects' actual performance is compared. This is reasonable, and will predict performance accurately, so long as the attributes actually used by subjects to describe the experimental stimuli approximately correspond to those assumed by the experimenter.
Figure 1 also illustrates another point, namely, that the interfeature correlations need not be perfect for categories to be distinguished based on these correlations (see Stimulus Set 4). In principle, a category would have positive utility so long as some of its features could be predicted with greater-than-chance reliability. This is consistent with the arguments of Wittgenstein (1953), Rosch (1975), and others that natural categories are not defined in terms of necessary and sufficient features, but rather are often characterized by probabilistic features and fuzzy boundaries. Furthermore, defining category structure in terms of predictive utility (i.e., feature correlations) is consistent with the functional role of categories in making predictions, drawing inferences, and completing patterns based on partial information (e.g., Clapper & Bower, 1991; Holland, Holyoak, Nisbett, & Thagard, 1986; Schank, 1982).2

Theories of Unsupervised Learning

We distinguish two general approaches to capturing correlational patterns, each of which has been implemented by several models in the empirical literature. The first approach is to represent feature correlation patterns directly, for example, within a correlational matrix, rather than partitioning the domain into separate categories. We will refer to this as the autocorrelation approach because models of this type assume that learners monitor the strengths of association (correlation) between individual pairs of features. The only learning mechanism required by this theory would be a process for modifying correlational associations or rules. Such associations would be strengthened by repetition and weakened by decay, interference processes, or both. If some features within a stimulus set were consistently correlated in their appearance, their strengths of association would increase relative to those of uncorrelated values. Given such a correlational record in memory, subjects could fill in missing features of an incomplete pattern, distinguish correlated from uncorrelated features, and perform other such inferences normally associated with category-level knowledge. It is also important to note that this inferential power could be gained without any explicit categorization of the stimulus set.

There are two general types of autocorrelation theories. The first assumes that correlational associations between all presented features are strengthened simultaneously on each trial (e.g., J. A. Anderson, 1977; J. A. Anderson, Silverstein, Ritz, & Jones, 1977; McClelland & Rumelhart, 1985; Rumelhart, Hinton, & McClelland, 1986). We can refer to these models as matrix autocorrelators because memory is viewed as a matrix of interfeature correlations that are continually updated by new experiences. A specific example of this class would be the one-layered autoassociator model of J. A. Anderson (1977). The second type of such autocorrelation theories are the rule-sampling or hypothesis-testing theories, in which correlational hypotheses are tested sequentially (usually one per trial) against the observed features in each instance (e.g., Billman & Heit, 1988; Davis, 1985). These rules are strengthened by confirmation and may be weakened by disconfirmation on a given trial. The main difference between these theories and the matrix models is in whether all the interfeature correlations provided by an instance are strengthened simultaneously or sequentially and how many interfeature correlations are updated on each trial.

The second approach to capturing correlational patterns in a stimulus domain is to explicitly partition that domain into separate categories and store information about each category in separate data structures (e.g., schemas or prototypes). Within this approach, which we refer to as category invention, feature correlations are represented indirectly by (a) partitioning stimuli into explicit subsets or categories in accordance with correlational patterns and (b) accumulating summary norms separately for each category. These summary norms contain information about the expected features of individual instances. If only stimuli that contain a particular pattern of correlated features are assigned to a given category, then norms computed across this selected subset of instances will capture their correlational structure.

The major issue for the learner, according to this theory, is determining when and on what basis to create new categories. In many statistical clustering models of category learning, it is assumed that the learner will first scan an entire set of stimuli before computing the optimal classification scheme for that set (e.g., Fried & Holyoak, 1984; Michalski & Stepp, 1983). This assumption is generally unrealistic for human learners. Because of the assumptions of this theory, people must examine stimuli one at a time and update relevant category norms in response to each. Rather than computing global classification schemes across whole stimulus sets, humans are more likely to be opportunistic categorizers, creating new categories as they are needed to accommodate novel stimuli that do not fit into existing categories (e.g., J. R. Anderson, 1991; Clapper & Bower, 1991; Holland et al., 1986; Schank, 1982). We refer to this as the incremental learning assumption.

Incremental learning implies that subjects attempt to categorize each presented stimulus and that summary knowledge about the category provides a framework within which instances are described and compared with normative expectations. Assuming that there is a good enough fit between an instance and a known category, the features of that instance

2 Note that this definition of categories in terms of correlational patterns does not imply that all members of a given category must be more similar to each other than to any nonmember. For example, in Figure 1, Stimulus Set 2, the instance 11122222, which is a member of Category A, is more similar to instance 22222222, which is a member of Category B, than to fellow Category A instance 11111111. We define categories in terms of predictive utility rather than in terms of similarity, or family resemblance (e.g., Rosch & Mervis, 1975). As shown by Figure 1, in some domains there may be no categorization scheme in which all members are more similar to each other than to any nonmembers. Nevertheless, there may be useful structure to be captured in such domains (i.e., correlational patterns among some attributes of the stimuli). This definition, however, does not exclude the possibility that the most natural categories, those that are easiest to learn and use, may have members that are highly similar to each other and dissimilar to members of other categories. Thus, our definition of categories does not contradict the arguments of Rosch and Mervis (1975) and others that so-called basic level categories tend to exhibit family resemblance structures.
will be used to update the norms of its category, that is, the expectedness or subjective probability of the presented attribute values will be incremented in the category norms. But in some cases, an instance may fit poorly into even the closest available category. For example, it may be describable in terms of the attributes associated with that category, but it may also violate several of its default expectations (see, e.g., Schank, 1982). In these cases, a new category could be created to accommodate that stimulus. When later instances are presented similar to that which triggered the new category, they will also be assigned to this category.

The norms for a given category might be represented in several ways, including prototypes, schemas, scripts, frames, various networks, and production rules (e.g., J. R. Anderson, 1991; Holland et al., 1986; Kahneman & Miller, 1986; Minsky, 1975; Rumelhart & Ortony, 1977; Schank, 1982; Schank & Abelson, 1977). All of these approaches are capable of representing statistical summaries of the properties of instances within a category, that is, of resembling a subjective probability distribution for the occurrence of different values of each attribute. For the present purposes, the differences between these various methods of representing category norms are relatively unimportant, and they are de-emphasized throughout this article. The major claims of the category invention approach pertain not to details of how category norms are represented in memory but rather to the explicit separation of norms from different categories on the basis of their perceived contrast and to the selective assimilation of instances to these categories. It is this discrete partitioning of experience that generates the major predictions that are tested here.

Category Learning in Discrimination Tasks

The first step in testing these theories is developing an appropriate task or paradigm in which unsupervised learning can be reliably observed and investigated. In the absence of such tasks, little prior empirical study of unsupervised category learning has occurred. Previous supervised classification experiments provide little guidance toward developing unsupervised learning tasks. Traditionally, supervised learning has been measured by classification accuracy, where subjects classify presented instances into alternative categories provided by the experimenter (e.g., Bruner et al., 1956). Because subjects in unsupervised learning are not given predefined categories, classification accuracy obviously cannot be used to measure learning in these tasks.

If categories are defined in terms of correlational patterns within a domain of stimuli, then acquisition of such categories would be implied whenever the subjects’ performance reveals their sensitivity to these patterns. One indication of such sensitivity would be if subjects in certain tasks responded differently to correlated attribute values than to uncorrelated values. One task we have investigated that has these properties is presented to subjects as an instance-discrimination (identification) task in which subjects are asked to learn to distinguish among a set of presented stimuli so that they can respond uniquely to each one. In learning to identify each individual instance of a set, subjects must first learn how that instance differs from the other stimuli presented during training. In other words, subjects must learn which features or combinations of features specify that instance’s unique identity and must exclude all possible lures within the presented stimulus set.

If the subject’s task is to memorize a collection of stimulus patterns, then their labor can be greatly reduced by noticing and taking advantage of redundancies among some of the features. These advantages can be illustrated with the task of memorizing the 16 stimulus patterns shown as rows in Stimulus Set 1 of Figure 1. Here Attributes 1 through 5 are redundant, with values of 1 in one cluster (Category A) and values of 2 in the other cluster (Category B). Although there are eight attributes, and potentially $2^8 = 256$ patterns in the sample space, the 16 patterns actually presented can be uniquely identified by their values on four different attributes—the last three and some (any) one of the first five. Rather than memorizing the configuration of eight bits per stimulus, the optimal learner could memorize the 16 stimuli by recording only four bits of information for each pattern—namely, the category (or any of the default values, each of which predicts the other four) and the values of the last three (unpredictable or nonredundant) attributes. It is also important to note that once the value of one of the default attributes is specified, the other four defaults are unnecessary for identifying a unique stimulus. This contrast suggests that subjects who have learned the subjective categories (or clusters of interfeature correlations) will treat default attributes differently from variable attributes as they try to memorize each instance.

The Attribute-Listing Task

The foregoing discussion suggests that if an observable index of feature weighting could be developed for instance-discrimination tasks, then such tasks might be used to investigate unsupervised learning. In the experiments described in Clapper and Bower (1991), subjects were presented with a series of instances and were asked to list those features that they considered most informative for distinguishing each instance from all those that they had seen previously in the series. Subjects were told to imagine that they would have to use their feature list at some later time to pick out the current instance from among a field of similar distractors in a multiple-choice recognition test. They were instructed to list only those features they would need to pick out the current stimulus in such a discrimination test and to omit features that they would not need even if these were physically very salient or prominent. As in Figure 1, the stimuli in these tasks were composed of several attributes, each with two or more alternative values. Categories in the stimulus sets were defined in terms of correlated attribute values.

Within the autocorrelational approach, the probability of listing a given attribute value should depend on how strongly it is correlated with other values of the current instance. Thus, learning in a given condition is defined as subjects’ sensitivity to differences in the degree of correlation among different pairs of attribute values, that is, sensitivity to the fact that some values of an instance are mutually redundant and others are not. This sensitivity is measured in terms of differences in
listing probability for correlated versus uncorrelated values (i.e., in terms of subjects’ observed preference for listing variables rather than defaults).

The interpretation of the listing task in terms of category invention is similar to its interpretation in terms of autocorrelation. Here, the probability of listing a given value should be a function of its expectedness or probability of occurrence within the current reference category. Learning is defined as sensitivity to differences in expectedness between variables and defaults, again measured by differences in their probability of listing (i.e., by subjects’ preference for listing variables over defaults).

By subtracting the proportion of defaults listed on a given trial from the proportion of variables listed, we may compute a quantitative index of learning for that trial. This index provides a way to compare the level of learning on different trials of an experiment; for example, if the preference measure is statistically greater on Trial $n + 1$ than on Trial $n$, then it can be inferred that some learning has occurred over that interval of trials. In experiments reported by Clapper and Bower (1991), such a bias in favor of listing uncorrelated variables evolved gradually over trials as successive instances were encountered and subjects learned their consistent properties.

**Distinguishing the Theories**

In the present article, we use this preference measure to compare learning under different experimental conditions, that is, to evaluate the effects of specific independent variables on unsupervised learning. To test the autocorrelation versus category invention theories described earlier, we looked for some variable that the theories would expect to have different effects on learning. We noted that the theories we considered differed in their predictions of how the particular sequencing of training instances from two categories would affect the rate at which categories are learned. Consequently, the experiments described in this article rely on such sequence manipulations to test the autocorrelation versus category invention theories.

We assume that learners update their category knowledge (by modifying existing categories or creating new ones) following the presentation of each new training instance. Given this incremental learning assumption, category invention should be highly sensitive to the order in which instances from different categories are presented during training. In particular, learning should be greater when categories are acquired one at a time (e.g., when Category A is well-learned prior to encountering any instances of Category B) than when instances of different categories are presented together from the start of training. In the latter (mixed) sequence, learners may simply lump both types of instances together into a single category, thus, failing to capture the correlational patterns in the stimulus set.

To understand these predicted sequence effects, imagine an experiment in which instances of Category A are presented for the first $n$ trials, followed by an instance of Category B on Trial $n + 1$. Given this arrangement, we would then ask how the probability of creating a new Category B on Trial $n + 1$ would vary as a function of $n$. To answer this question, consider that any reasonable function for inducing category norms from a series of training instances should show some sensitivity to basic statistical parameters (e.g., sample size and variability) that greatly affect the reliability of its norms (generalizations). For example, people should be more confident in assigning grey as the default color of elephants after they have seen many elephants, all of which were grey, than if they have seen only one elephant, which happened to be grey. Applying this observation to the experimental situation described above, as successive instances of Category A are presented (i.e., as the value of $n$ is increased), one can see that the consistent default attributes of that category should increase in their expectedness. As the learner’s confidence in the Category A norms increases, so should the perceived contrast between these norms and the first instance of Category B, which violates several default values of Category A. Thus, the probability of creating a new category in response to the first instance of Category B should increase with the number ($n$) of prior instances of Category A.

This analysis implies that presenting the first instance of Category B following only a few instances of Category A should lead to a higher probability that the two types of instances will be assimilated to a single, overarching category. This would occur because the features of the Category B instance would be compared with a relatively weak set of norms for Category A; hence, the perceived contrast between these norms and the instance of Category B would be reduced. If both types of instances were assimilated to a single category, the learner would then simply average over the feature correlations within the A and B categories, so that the correlational information conditional upon the two categories would be lost. Because neither instances in Category A nor instances in Category B would contrast strongly with this aggregated category on subsequent trials (assuming both were presented in random order), subjects might have difficulty unlearning these overgeneralized norms and discovering the correct category-level discriminations.

Autocorrelation models do not possess the same inherent tendency toward sequence sensitivity shown by category invention models when incremental learning is assumed. For example, it is easy to imagine a basic autocorrelation model that simply adds to incremental frequency counts within a correlational matrix each time a new instance is encountered. In principle, such a model would be completely immune to sequence effects on final learning (i.e., the final count in the matrix would be the same regardless of the order in which instances were presented). Thus, the model suggests that learning that the presence of large wings predicts black eyes in some insects would not affect learning that in other insects the presence of small wings predicts white eyes.

Although sequence sensitivity is not implied by the autocorrelational approach, it is important to ask whether it is possible to develop plausible models within this approach that mimic the particular type of sequence sensitivity expected by category invention. Existing autocorrelation models do not display sequencing effects similar to those of category invention. For example, autocorrelation models developed within the connectionist framework generally predict sequence effects that are almost the opposite of those expected by category invention.
models (e.g., J. A. Anderson, 1977; Rumelhart et al., 1986). These models predict that correlational learning should be improved if instances of both categories are presented mixed together (e.g., in random alternation) from the beginning of training. Presenting a block of instances from Category B following an earlier block of instances from Category A causes massive forgetting of correlational associations learned during the Category A block, a phenomenon referred to as catastrophic interference (McCloskey & Cohen, 1989; Ratcliff, 1990). By contrast, category invention theory predicts better learning in a blocked condition than when instances are presented in a mixed sequence and expects no catastrophic interference between categories.3

Models of unsupervised learning based on serial hypothesis testing (e.g., Billman & Heit, 1988; Davis, 1985) also fail to reproduce the sequence effects expected by category invention. In such models, there is little reason to expect interference between different correlational patterns within either blocked or mixed sequences. For example, the rule “large wings implies black eyes” neither confirms nor disconfirms the rule “small wings implies white eyes,” and there is no obvious reason why learning one should increase the difficulty of learning the other. Indeed, the focused-sampling assumptions of Billman and Heit (1988) seem more compatible with positive transfer across categories (at least if the categories differ by contrasting defaults on the same set of attributes).

To reproduce the sequence effects predicted by category invention, autocorrelation models would have to include a process that strongly reduces correlational learning when instances of two patterns are mixed together, but not when they are presented in separate blocks. For example, an autocorrelation model could assume that correlational learning is subject to associative interference, or fan effects, similar to those studied in experiments on paired-associate learning (e.g., Postman, 1971) and sentence memory (e.g., J. B. Anderson, 1976; 1983). Thus, learning an association between a particular pair of attribute values (e.g., large wings with black eyes) might interfere with learning associations between other values of the same attributes (e.g., small wings with white eyes). This autocorrelation-with-interference theory could accommodate some of the results predicted by the alternative, category invention theory. For example, the interference theory predicts that correlations within Category A would be learned more slowly if instances of Category A were interwoven in the training sequence with instances of Category B than if the instances of Category A were presented alone or prior to any instances of Category B.

The problem with such interference theories is that interference should occur between different correlational patterns regardless of whether training instances are presented in a blocked or mixed sequence, whereas category invention predicts interference only in mixed sequences. Thus, an interference theory expects prior learning of instances of Category A to impair correlational learning in a later block of instances of Category B (similar to the negative transfer between lists observed in many paired-associate learning experiments, e.g., Postman, 1971). The more instances of Category A that are presented prior to Category B, the greater the negative transfer and the slower should be the learning of Category B correlations. This contradicts the prediction of category invention that presenting more instances of Category A prior to Category B should increase the probability of creating a separate Category B and thus improve learning of both Category A and B defaults.

Such correlational interference would also imply that learning Category B in a blocked sequence would cause retroactive interference and reduce prior learning of Category A (e.g., Postman, 1971), although this effect need not be as strong as the catastrophic interference predicted by connectionist autoassociators. (At least, catastrophic retroactive interference is not generally observed in standard experiments on associative interference.) As noted previously, the category invention theory expects no interference across categories once separate categories have been formed.

In summary, such variations of the autocorrelation approach appear unable to mimic the particular pattern of sequence sensitivity expected by category invention theories. Thus, demonstrating superior learning and a lack of interference between categories in blocked training sequences would provide evidence for a nonincremental, contrast-based process of category invention.

Experiment 1

The aim of this experiment was to evaluate the attribute-listing task as an index of unsupervised learning and to test the predictions of the two theories concerning sequence effects. Subjects’ listing of attributes was compared in three conditions. In the blocked condition, the stimuli were partitioned into two categories based on patterns of correlated attribute values. The training instances were blocked by categories (i.e., a series of instances from one category was presented followed by a series of instances from the other category). Following these two training blocks was a test, or transfer, block in which several instances of both categories were presented in random order. In the mixed condition, the same instances were presented as in the blocked condition, but instances of both categories were randomly interspersed in the training sequence rather than being grouped into separate blocks. In the control condition, all the attributes of the stimuli varied independently, so that none of the attributes were correlated and the stimulus set was not partitioned into distinct categories. The same final test block that was presented in the

3 Most connectionist models that could be applied to unsupervised learning are apparently subject to catastrophic interference, even when these models are not strict autocorrelators (but see Carpenter & Grossberg, 1987). The reason is that such models encode knowledge about contrasting categories as patterns of activation over the same set of network units even when these models do contain an explicit category level of representation (e.g., output units corresponding to different response categories). Because most connectionist models do not separate knowledge about different categories in memory the way that prototype or schema models do, different patterns are liable to interfere with each other, especially when they are learned separately, for example, in blocked training sequences (McCloskey & Cohen, 1989).
blocked condition was also given in the mixed and control conditions.

The first two conditions provided a test of the two models of unsupervised learning described earlier. Category invention theory implies that early aggregation may occur when contrasting categories are presented in a mixed sequence, and so poorer learning was predicted in the mixed condition than in the blocked condition. An autocorrelation model could accommodate interference between categories in the mixed condition by assuming that associative interference results from learning correlations among different values of the same set of attributes. However, this leads to the prediction that interference should be observed between the categories in the blocked condition as well as in the mixed condition, as noted earlier. Specifically, the autocorrelation-plus-interference hypothesis predicts (a) that the second category in the blocked condition should be learned more slowly than the first because of proactive interference or negative transfer from the first category, and (b) that once this second category is learned, it should produce retroactive interference on subjects' memory for the first category, that is, that evidence of forgetting or unlearning should be obtained when instances of the first category are presented in the final test block. By contrast, the category invention theory expects little interference of any kind in the blocked condition.

The third condition was included in this experiment as a control to evaluate learning in the other two conditions. This condition was identical to the others except that the stimuli lacked correlated attributes. Thus, any differences in performance between this condition and the correlated-attribute conditions would be due to these correlations rather than to other, extraneous, factors.

**Method**

*Subjects.* The subjects were 30 Stanford University undergraduates participating in partial fulfillment of an introductory psychology course requirement.

*Procedure.* Subjects were tested in groups of 8 to 10 for a single session of 40 to 50 min. The training instances were realistic line drawings of fictitious insects (see Figure 2) presented in a 42-page booklet that measured 8 in. by 11.5 in. (20.3 cm by 29.2 cm). The first two pages of this booklet contained full instructions and an agreement that subjects signed to indicate their informed consent to participate. A single training instance (insect picture) appeared on each subsequent page, together with brief instructions for the experimental task.

Subjects were instructed to write on each page the “distinctive” properties of each individual insect, where distinctive properties were those that would be useful for distinguishing the current instance from others of the same general type. Subjects were told to imagine that they were writing their lists in order to prepare for a later multiple-choice recognition test in which they would have to match up each list with the correct insect from among a large number of distractor items (i.e., other bugs from the same test booklet). Subjects were instructed to list only those properties that would be useful for identifying an insect on such a test and to omit nondistinguishing properties even if they were highly prominent or noticeable. They were further told to look only at the page of the booklet that they were currently working on and not to look backward or forward at other pages.

Subjects were allowed to complete the experimental task at their own pace. Once they had finished, they were given a debriefing page that explained the procedures and goals of the experiment and were allowed to leave.

*Materials.* The stimuli were line drawings of fictitious insects, all of which shared a common base structure (e.g., head, thorax, abdomen) plus eight dimensions of variation (attributes), such as wing shape, abdominal markings, eye color, and so forth (see Figure 2). Each attribute had either two or four discrete values (e.g., different wing shapes, differently colored eyes) depending on the experimental condition to which it was assigned.

The stimuli shown to a given subject were constructed according to one of two different plans depending on a subject’s assigned condition (see Table 1). In two of the three experimental groups, the stimulus set was partitioned into two distinct categories defined by contrasting sets of correlated attribute values. In these correlated groups, five of the eight attributes were binary (two-valued), and their values were perfectly correlated across the instances such that each instance contained one of two possible sets of correlated values (denoted as Values 1 or 2 in Table 1). An instance’s category membership was defined by which of these two clusters of correlated values it contained. These values are referred to as the *default* values of each category.
instances in this test block was randomized (the same randomizations category were presented together in a mixed sequence. The order of instances was presented twice). Following this training phase was a instances from the two categories were presented. Used the same stimuli and differed only in the order in which training subjects conditions, two of which had correlated values and one of greater in the correlated groups than among the controls, this conditions. But if the preference for listing variables over defaults is defaults should be observed in the control group as in the correlated groups (see Table 1). The stimuli in the uncorrelated group were equivalent to those in the correlated condition, instances were presented in random order, with the restriction that no more than three instances from the same category could occur in a row. In the control condition, instances were presented in the same order in both conditions.

The remaining three attributes in the correlated conditions had four values and were variable within each category. Two of the four values occurred with equal probability in instances of Category A, whereas the other two occurred with equal probability in instances of Category B. These attributes were uncorrelated within each category (i.e., they varied independently across instances of that category). Within these constraints, \(2^3 = 8\) instances were generated from each category, for a total of 16 overall.

The stimuli in the control condition were equivalent to those in the two correlated conditions in the number of values assigned to each attribute (two or four), but these insects lacked correlated attributes present in the other conditions. Two attributes were correlated in all conditions; these were the wing shape and body shape attributes, which we judged to be the most salient attributes of the insects. These defaults, which were constant across all three groups, are referred to as base defaults. The four-valued variables were coordinated with the base defaults in the same way in the uncorrelated group as in the correlated groups (see Table 1). The stimuli in the uncorrelated groups can be divided into two pseudocategories on the basis of the base defaults and the pattern of dependent variation of the four-valued variables. However, three binary attributes that had correlated defaults in the other conditions occurred as uncorrelated variables in this condition.

The control condition was designed to show that any greater listing of variables over defaults in the correlated conditions could not simply be explained as an artifact due to variables possessing more possible values than defaults (four versus two). If this artificial explanation is correct, then the same degree of bias in reporting variables over defaults should be observed in the control group as in the correlated conditions. But if the preference for listing variables over defaults is greater in the correlated groups than among the controls, this difference must be due to subjects’ explicit or implicit correlational learning.

**Design.** The experimental design contained three between-subjects conditions, two of which had correlated values and one of which did not, as explained previously. The two correlated conditions used the same stimuli and differed only in the order in which training instances from the two categories were presented.

In the blocked condition, instances of Category A were presented in random order for the first 16 trials, followed by 16 trials in which instances of Category B were presented (each instance of the two categories was presented twice). Following this training phase was a final test block of eight trials in which four instances from each category were presented together in a mixed sequence. The order of instances in this test block was randomized (the same randomizations were used for subjects in all three groups), with the restriction that no more than two instances from the same category could occur in a row.

In the mixed condition, the same instances were presented as in the blocked condition, but in a different order. During the training phase, 16 instances from Category A and the 16 instances from Category B were presented in an intermixed sequence rather than blocked as in the previous condition. Instances from the two categories were presented in random order, with the restriction that no more than three instances from the same category could occur consecutively. A final mixed test block of eight instances from the two categories was then presented, the same as that used in the blocked condition, (i.e., the same specific insect pictures were presented in the same order in both conditions).

In the control condition, instances were presented in random order for the first 32 trials, except that no more than three instances with the same base default values were allowed to occur in a row during this phase. The final eight test trials were identical to those of the category conditions (i.e., five attributes were correlated during this block).

**Counterbalancing the design.** To construct stimuli from the specifications shown in Table 1, we first assigned particular stimulus attributes to abstract roles in the design. This assignment was held constant across all groups. With the exception of base defaults, each attribute had four values in half of the groups and two values in the other half of the groups. Two different stimulus sets were constructed for each of the three between-subjects conditions (blocked, mixed, and control); that is, six booklets were constructed and presented to different subjects. Attributes that were four-valued variables in one group were two-valued defaults in the other group from the same condition. This ensured that any effects due to materials (e.g., differences in the baseline salience or prominence of different attributes) would be balanced over the experiment as a whole.

**Results and Discussion**

Subjects’ attribute lists were coded in terms of whether or not each of the eight relevant attribute dimensions was mentioned on a given trial. 4 The main index of learning was the

4 Because this was a free-listing task, subjects generated their own response categories. For example, subjects shown a bug with large mandibles might describe the instance as possessing “big pincers,” “large mouthparts,” “oversized mandibles,” or a variety of other labels. Although the specific labels might vary among different subjects, it was generally clear which attribute was being referred to at
proportion of variables listed minus the proportion of defaults listed on a given trial. This difference is referred to as the 
preference score for each trial because it reflects subjects' preference for listing variables over defaults. Preference scores for all three conditions are shown in Figure 3.

We also recorded the proportion of base defaults listed on each trial, but because they were correlated in all groups and hence potentially contaminated by materials effects, subjects' listing of these attributes did not provide the same unambiguous measure of learning as did the preference scores. Hence, we focus on preference scores as the principle dependent measure in most of the following discussion.

Examination of Figure 3 reveals, first, that preference scores were higher overall in the two correlated conditions (blocked and mixed) than in the control condition. Averaged over all trials, four-valued attributes were listed 19.6% more often than two-valued attributes in the control condition, a significant preference, t(9) = 3.93, SE = 0.05, p < .01. However, this preference was much stronger in the other two groups. Variable attributes were listed 74% more often than defaults in the blocked condition, t(9) = 9.65, SE = 0.077, p < .001. This preference was significantly greater than the corresponding difference in the control condition, t(18) = 5.95, SE = 0.092, p < .001. In the mixed condition, variables were listed about 55% more often than defaults, t(9) = 10.38, SE = 0.053, p < .001; this effect was also significantly larger than the 19.6% preference in the control condition, t(18) = 4.84, SE = 0.073, p < .001.

The contrasting results for the correlated versus control conditions indicate that the preference for listing variables over defaults in the correlated conditions was in large part due to the correlations themselves, not simply to the fact that uncorrelated attributes had a larger number of possible values than correlated attributes. Thus, subjects in the correlated groups must have internalized the correlational structure of the stimulus set in some manner, either by tracking pairwise correlations or by partitioning the set into separate categories.

The category invention theory predicts that preference scores would show rapid learning of both categories in the blocked condition but that learning in the mixed condition would be slower. The data are generally consistent with this prediction. If one examines the data plotted in Figure 3, it is apparent that the preference for studying variables over defaults increased rapidly for both categories in the blocked condition. Preference scores increased from 0.03 on the first Category A trial to .90 on the eighth and remained fairly stable thereafter; the linear trend over the first eight trials was highly significant, t(9) = 9.01, SE = 0.59, p < .001. Subjects sharply increased their listing of defaults when the first instance of Category B was presented. The resulting decrease in preference scores, compared with the immediately preceding Category A trial, was highly significant, t(9) = 6.31, SE = 0.122, p < .001. Thereafter, preference scores increased rapidly from .17 on this first Category B trial to a maximum of .83 by the sixth. The linear contrast over the first half of this block was statistically significant, t(9) = 4.58, SE = 0.63, p < .01. However, no significant change occurred over the remaining nine instances in this block.

Recall that the autocorrelation-with-interference hypothesis predicts that prior learning of Category A should reduce subsequent learning of Category B because of negative transfer or proactive interference effects. But no such interference occurred in the present experiment. Learning of Category B appeared to occur at least as rapidly as that of Category A, and there was no significant difference between asymptotic learning of the two categories (i.e., when preference scores averaged over the last eight instances of each were compared), t(9) = 1.54, SE = 0.065, p > .10. This absence of proactive
interference appears to be a strike against the autocorrelation theory but is consistent with category invention models.

Preference scores during the mixed test block did not differ significantly from those of the earlier blocks. This was true when the test block was compared to the last eight instances of Category A, \( t(9) = 1.64, SE = 0.067, p > .10 \), as well as to the last eight instances of Category B, \( t(9) = 0.35, SE = 0.022, p > .50 \). In addition, preference scores during the test block did not differ between the two categories, \( t(9) = 0.16, SE = 0.504, p > .50 \). These results indicate that the learning observed during the earlier training blocks, in which instances of the same category were presented for many trials in succession, generalized to a different context in which the two categories were mixed. In other words, learning was stable over changes in the learning environment (Carpenter & Grossberg, 1987). There was no evidence for retroactive interference from learning Category B upon test performance on Category A, as would have been expected in an autocorrelation-with-interference framework.

Learning occurred somewhat more slowly in the mixed than the blocked condition. Preference scores increased over the entire training block for each category. This increase was significant for both Category A, \( t(9) = 4.18, SE = 3.55, p < .01 \) and Category B, \( t(9) = 3.74, SE = 3.31, p < .01 \). However, preference scores were greater in the blocked than the mixed condition over the first eight instances shown of Category A, \( t(18) = 3.69, SE = 0.079, p < .01 \) and of Category B \( t(18) = 2.34, SE = 0.12, p < .05 \). The same comparison was marginally significant over the second eight instances of Category A, \( t(18) = 1.96, SE = 0.10, p < .10 \). Pooled over all 32 training trials, preference scores were significantly higher in the blocked than the mixed condition, \( t(18) = 2.46, SE = 0.089, p < .05 \). However, the blocked and mixed conditions did not differ significantly during the test block, \( t(18) = 0.71, SE = 0.14, p > .25 \). This suggests that although learning occurred more rapidly in the blocked condition, subjects in the mixed condition were able to catch up by the end of training.

The faster learning that was due to category blocking is consistent with the category invention theory because it expects subjects to have difficulty separating categories presented in a mixed sequence. As noted earlier, however, an autocorrelation model could explain negative transfer in the mixed condition as being due to interference or unlearning of correlations among different feature pairs. However, such an interference process predicts a different pattern of results in the blocked condition than was shown by these data. First, it implies that prior learning of Category A should interfere with subsequent learning of Category B. However, these data show no such negative transfer; the second category was learned at least as fast as the first in this group. Second, an autocorrelation-with-interference model also predicts that Category B should exert strong retroactive interference on Category A in the blocked condition. As noted earlier, no evidence of such interference was obtained in the final test trials of this experiment. This lack of retroactive interference is particularly embarrassing for connectionist autocorrelators, which predict catastrophic interference from learning the Category B correlations on subjects’ memory for the earlier Category A correlations (McCloskey & Cohen, 1989; Ratcliff, 1990).

Although the preference scores showed no evidence of retroactive interference during the test block, there was some evidence that presenting instances of the two categories in a mixed sequence increased the salience of their category membership. Recall that base defaults were the most physically prominent attributes of the insect stimuli, and it was considered likely that subjects would tend to list these particular attributes when they wished to indicate an instance’s category membership. Although caution must be exercised when interpreting listing patterns for base defaults, because these attributes were correlated in all groups and hence their data may be contaminated with unbalanced materials effects, it appears that base defaults were often used by subjects to indicate the categorization of each instance. Consistent with this explanation, higher listings were observed for base defaults in the mixed test block of the blocked condition, in which the categorization of instances varied from trial to trial, than in the last eight trials of the preceding same-category training blocks, in which categories were constant and could be inferred from local context, \( t(9) = 2.48, SE = 0.081, p < .05 \). No such increase occurred for either variables, \( t(9) = 1.00, SE = 0.004, p > .25 \), or for regular defaults, \( t(9) = 1.54, SE = 0.035, p > .10 \). In other respects, the base defaults behaved like the regular defaults in the blocked condition, decreasing strongly during the first six instances of each category: \( t(9) = 2.83, SE = 0.478, p < .05 \) for Category A, and \( t(9) = 6.85, SE = 0.255, p < .001 \) for Category B.

By contrast, base defaults remained fairly constant throughout the experiment in the mixed condition, showing no significant decreasing trends and remaining significantly higher than the regular defaults, \( t(9) = 2.55, SE = 0.067, p < .05 \). Subjects who learned the categories in the mixed condition would have needed to explicitly indicate the category membership of each instance throughout the experiment because this could not be inferred from context. To do so, they should have continued listing at least one of the base defaults as shown by the present data.

Experiment 2

The aim of this experiment was to extend the results of Experiment 1 by testing further predictions of the category invention theory. Subjects were randomly assigned to two conditions. In the contrast condition, a pretraining block of 8 instances of Category A was followed by a test block of 12 instances of Category A and 12 instances of Category B that were presented in mixed sequence. In this condition, subjects should learn strong Category A defaults prior to encountering their first instance of Category B. They should readily notice the contrast between the two categories when they encounter this instance of Category B and rapidly learn the default values of the newly invented Category B without unlearning or weakening the prior Category A norms.

In the second, practice, condition, a mixed pretraining block of four instances of Category A and four instances of Category B was followed by the same test block as in the contrast condition. Category invention implies that subjects may aggregate the two types of instances into a single category, thereby pooling and obscuring the correlational structure of the
stimulus set. The result would be reduced learning of both categories in this condition. By contrast, the autocorrelation theory expects better learning of Category B in the practice condition because correlational associations among Category B defaults would receive more practice (repetitions across different instances) in that condition. (A total of four instances of Category B were presented during the pretraining block in the practice condition, whereas no instances of Category B occurred prior to the test block in the contrast condition.)

The theories also make different predictions about transfer of learning from one category to the other. First, increasing the number of instances of Category A, from four in the practice condition to eight in the contrast condition, is expected by category invention theorists to impair learning of Category B. This would seem to be an example of positive transfer from Category A to Category B. Second, increasing the number of instances of Category B, from zero in the contrast condition to four in the practice condition, is expected by category invention theorists to impair learning of Category A, which is an example of interference or negative transfer from Category B to Category A. This seeming paradox—positive transfer from A to B combined with negative transfer from B to A—makes sense in terms of category invention because this theory assumes that the particular sequence in which instances are presented affects the probability of creating separate categories by either highlighting or camouflaging the differences between them. By contrast, the predicted interaction of transfer and repetition effects with the sequencing and number of instances from each category makes little sense within the autocorrelational framework. If the predicted pattern of results obtains, it would provide strong evidence for the existence of a category invention process in unsupervised learning.

Method

Subjects. The subjects were 40 undergraduate students of San Jose State University participating in partial fulfillment of an introductory psychology course requirement.5

Procedure. Subjects were tested in groups for a single session of 30–45 min. The training instances were line drawings of fictitious insects presented in booklets similar to those used in Experiment 1. The attribute-listing procedure was identical to that of Experiment 1, except that the present experiment consisted of 32 instead of 40 trials.

Materials. The same type of pictorial insect stimuli as in Experiment 1 were used. These stimuli all shared a common base structure (e.g., head, thorax, abdomen) plus eight dimensions of variation (attributes), such as wing shape, abdominal markings, eye color, and so forth. Five of the eight attributes had two values, and these values were correlated across instances such that the set was partitioned into two distinct categories defined by contrasting sets of default attribute values (see Table 2).

The remaining three attributes had four values, two of which occurred with equal probability in Category A and the other two of which occurred with equal probability in instances of Category B. These variable attributes were uncorrelated within each category (i.e., they varied independently across instances of that category). A total of eight instances (23) could be generated within each of the two categories within these constraints. All 16 possible instances were presented to subjects in this experiment.

Design. Two between-subjects conditions were tested in this experiment. In the contrast condition, only instances of Category A were presented for the first eight trials, followed by a mixed block of 12 instances of Category A and 12 instances of Category B. The first block of eight trials was referred to as the pretraining block, whereas the second block of 24 trials was referred to as the test block. The first instance of the test block was always a member of Category B. Instances of both categories were thereafter presented in a randomly ordered, intermixed sequence, with the constraint that no more than three instances from the same category be allowed to appear in succession.

In the practice condition, the eight instances from the pretraining block consisted of four from Category A and four from Category B, rather than eight from Category A as before. The four instances from each category were selected so that both values of each variable attribute occurred twice, and none of the variable attributes was correlated with any of the others. These instances were presented in a random order, with the restrictions that the first instance be a member of Category A and that no more than two instances from the same category occur in sequence. The same 24-instance test block was used as in the contrast condition. Note that the only difference between the two conditions is that in the practice condition, four instances of Category B were substituted for the four instances of Category A that were presented in the contrast condition.

Counterbalancing the design. The counterbalancing scheme for this experiment is illustrated in Table 2. All of the attributes had four values in one condition and two (correlated) values in the other, except for the first two attributes. The first two attributes were base defaults, which consisted of the wing shape and body shape attributes, as in Experiment 1. These were two-valued and correlated in both conditions. The balancing scheme shown in Table 2 ensured that materials effects (e.g., differences in baseline prominence of different attributes) would be balanced over the six attributes that were not base defaults. Half of the subjects in the contrast and practice conditions were tested with Stimulus Set 1 and half with Stimulus Set 2.

Results and Discussion

The same attribute-listing data as in Experiment 1 was collected in this experiment. The preference scores (listing of

5 Thanks to Forest Jourden and to the San Jose State University Psychology Department for facilitating access to their subject pool for this experiment.
As shown in Figure 4, learning was higher in the contrast than in the practice condition throughout the experiment. Preference scores increased significantly during the pretraining block in both contrast, \( t(16) = 5.23, SE = 0.774, p < .001 \), and practice conditions, \( t(17) = 4.86, SE = 0.627, p < .001 \); however, preference was higher overall in the blocked condition, \( t(33) = 4.83, SE = 0.707, p < .001 \). This result is not surprising because subjects in the mixed condition were shown instances of two categories during pretraining while those in the blocked condition only had one category to learn during this interval.

Turning to the test block (the numbered trials in Figure 4), one can see that learning of both categories was higher in the contrast condition than in the practice condition. Preference scores for Category A showed a significant decrease on the first Category A instance of the test block, relative to the last instance of the pretraining block, \( t(16) = 2.85, SE = 0.075, p < .02 \). Thus, encountering the first instance of Category B at the beginning of the test block appeared to have a significant effect on Category A norms in this group. After this initial decrease, preference scores for Category A showed a modest but statistically significant increase over the remaining trials of the test block, \( t(16) = 2.36, SE = 0.959, p < .05 \). By contrast, preference scores for Category A in the practice condition showed neither the initial decrease, \( t(17) = 0.96, SE = 0.096, p > .50 \), nor the subsequent increasing trend, \( t(17) = -0.78, SE = 0.762, p > .20 \), observed in the contrast condition. Overall, preference scores for Category A during the test block were higher in the contrast condition than in the practice condition, \( t(33) = 3.40, SE = 0.114, p < .01 \).

Note that both the category invention and autocorrelational approaches can accommodate the finding that overall learning of Category A was greater in the contrast than in the practice condition. Initially lumping it with Category B, as occurred in the practice condition. Autocorrelation theory would expect better learning of Category A in the contrast condition because a larger number of instances in that category were presented in that condition.

Although both theories predict faster learning of Category A in the contrast condition, the autocorrelation approach has difficulty accommodating the detailed pattern of results from this condition. Thus, autocorrelation seems to imply that learning of Category A should have continued to increase following the pretraining block in the practice condition. Although Category A learning in the practice condition would have been expected to lag a few trials behind that in the contrast condition, in principle, asymptotic learning should have been about the same in both groups. However, Category A preference scores did not increase further during the test block of the practice condition; Category A learning appears to have stopped by the end of pretraining and never to have approached the asymptotic level attained in the contrast condition.

Category invention theory predicts that subjects should discriminate between contrasting categories better when one of the categories is learned first (contrast condition) than when instances of both are presented together from the start of training (practice condition). It is important to note that this result was predicted not only for the pretrained category (A) but also for the nonpretrained category (B). The better initial learning of Category A in the contrast condition was expected initially lumping it with Category B, as occurred in the practice condition. Autocorrelation theory would expect better learning of Category A in the contrast condition because a larger number of instances in that category were presented in that condition.

Although both theories predict faster learning of Category A in the contrast condition, the autocorrelation approach has difficulty accommodating the detailed pattern of results from this condition. Thus, autocorrelation seems to imply that learning of Category A should have continued to increase following the pretraining block in the practice condition. Although Category A learning in the practice condition would have been expected to lag a few trials behind that in the contrast condition, in principle, asymptotic learning should have been about the same in both groups. However, Category A preference scores did not increase further during the test block of the practice condition; Category A learning appears to have stopped by the end of pretraining and never to have approached the asymptotic level attained in the contrast condition.

Category invention theory predicts that subjects should discriminate between contrasting categories better when one of the categories is learned first (contrast condition) than when instances of both are presented together from the start of training (practice condition). It is important to note that this result was predicted not only for the pretrained category (A) but also for the nonpretrained category (B). The better initial learning of Category A in the contrast condition was expected initially lumping it with Category B, as occurred in the practice condition. Autocorrelation theory would expect better learning of Category A in the contrast condition because a larger number of instances in that category were presented in that condition.

Although both theories predict faster learning of Category A in the contrast condition, the autocorrelation approach has difficulty accommodating the detailed pattern of results from this condition. Thus, autocorrelation seems to imply that learning of Category A should have continued to increase following the pretraining block in the practice condition. Although Category A learning in the practice condition would have been expected to lag a few trials behind that in the contrast condition, in principle, asymptotic learning should have been about the same in both groups. However, Category A preference scores did not increase further during the test block of the practice condition; Category A learning appears to have stopped by the end of pretraining and never to have approached the asymptotic level attained in the contrast condition.

Category invention theory predicts that subjects should discriminate between contrasting categories better when one of the categories is learned first (contrast condition) than when instances of both are presented together from the start of training (practice condition). It is important to note that this result was predicted not only for the pretrained category (A) but also for the nonpretrained category (B). The better initial learning of Category A in the contrast condition was expected initially lumping it with Category B, as occurred in the practice condition. Autocorrelation theory would expect better learning of Category A in the contrast condition because a larger number of instances in that category were presented in that condition.

Although both theories predict faster learning of Category A in the contrast condition, the autocorrelation approach has difficulty accommodating the detailed pattern of results from this condition. Thus, autocorrelation seems to imply that learning of Category A should have continued to increase following the pretraining block in the practice condition. Although Category A learning in the practice condition would have been expected to lag a few trials behind that in the contrast condition, in principle, asymptotic learning should have been about the same in both groups. However, Category A preference scores did not increase further during the test block of the practice condition; Category A learning appears to have stopped by the end of pretraining and never to have approached the asymptotic level attained in the contrast condition.

Category invention theory predicts that subjects should discriminate between contrasting categories better when one of the categories is learned first (contrast condition) than when instances of both are presented together from the start of training (practice condition). It is important to note that this result was predicted not only for the pretrained category (A) but also for the nonpretrained category (B). The better initial learning of Category A in the contrast condition was expected initially lumping it with Category B, as occurred in the practice condition. Autocorrelation theory would expect better learning of Category A in the contrast condition because a larger number of instances in that category were presented in that condition.

Although both theories predict faster learning of Category A in the contrast condition, the autocorrelation approach has difficulty accommodating the detailed pattern of results from this condition. Thus, autocorrelation seems to imply that learning of Category A should have continued to increase following the pretraining block in the practice condition. Although Category A learning in the practice condition would have been expected to lag a few trials behind that in the contrast condition, in principle, asymptotic learning should have been about the same in both groups. However, Category A preference scores did not increase further during the test block of the practice condition; Category A learning appears to have stopped by the end of pretraining and never to have approached the asymptotic level attained in the contrast condition.

Category invention theory predicts that subjects should discriminate between contrasting categories better when one of the categories is learned first (contrast condition) than when instances of both are presented together from the start of training (practice condition). It is important to note that this result was predicted not only for the pretrained category (A) but also for the nonpretrained category (B). The better initial learning of Category A in the contrast condition was expected initially lumping it with Category B, as occurred in the practice condition. Autocorrelation theory would expect better learning of Category A in the contrast condition because a larger number of instances in that category were presented in that condition.

Although both theories predict faster learning of Category A in the contrast condition, the autocorrelation approach has difficulty accommodating the detailed pattern of results from this condition. Thus, autocorrelation seems to imply that learning of Category A should have continued to increase following the pretraining block in the practice condition. Although Category A learning in the practice condition would have been expected to lag a few trials behind that in the contrast condition, in principle, asymptotic learning should have been about the same in both groups. However, Category A preference scores did not increase further during the test block of the practice condition; Category A learning appears to have stopped by the end of pretraining and never to have approached the asymptotic level attained in the contrast condition.

Category invention theory predicts that subjects should discriminate between contrasting categories better when one of the categories is learned first (contrast condition) than when instances of both are presented together from the start of training (practice condition). It is important to note that this result was predicted not only for the pretrained category (A) but also for the nonpretrained category (B). The better initial learning of Category A in the contrast condition was expected initially lumping it with Category B, as occurred in the practice condition. Autocorrelation theory would expect better learning of Category A in the contrast condition because a larger number of instances in that category were presented in that condition.

Although both theories predict faster learning of Category A in the contrast condition, the autocorrelation approach has difficulty accommodating the detailed pattern of results from this condition. Thus, autocorrelation seems to imply that learning of Category A should have continued to increase following the pretraining block in the practice condition. Although Category A learning in the practice condition would have been expected to lag a few trials behind that in the contrast condition, in principle, asymptotic learning should have been about the same in both groups. However, Category A preference scores did not increase further during the test block of the practice condition; Category A learning appears to have stopped by the end of pretraining and never to have approached the asymptotic level attained in the contrast condition.

Category invention theory predicts that subjects should discriminate between contrasting categories better when one of the categories is learned first (contrast condition) than when instances of both are presented together from the start of training (practice condition). It is important to note that this result was predicted not only for the pretrained category (A) but also for the nonpretrained category (B). The better initial learning of Category A in the contrast condition was expected initially lumping it with Category B, as occurred in the practice condition. Autocorrelation theory would expect better learning of Category A in the contrast condition because a larger number of instances in that category were presented in that condition.
to increase the perceived contrast between the Category A norms and the features of the first instance of Category B, thereby increasing the probability that a new category would be created to describe the instance of Category B. Thus, better learning of Category B was predicted to occur in the contrast condition despite the larger number of instances presented to subjects in the practice condition (i.e., in spite of the fact that the interfeature correlations of Category B would have been repeated for a larger number of trials in the practice condition).

Consistent with this prediction, Category B was learned significantly better in the contrast condition than in the practice condition. Preference scores for this category increased quite rapidly in the contrast condition; the linear trend computed over the 12 trials of the test block was significant, \( t(16) = 4.36, SE = 1.210, p < .001 \). The linear contrast over the first 12 instances of Category B of the practice condition (4 of which were in the pretraining block) also showed a significant increase, \( t(17) = 2.63, SE = 1.269, p < .02 \). However, overall learning of Category B was higher in the contrast condition than in the practice condition. Averaged over the 12 test trials, preference scores in the contrast condition were significantly higher than those in the practice condition, \( t(33) = 2.09, SE = 0.109, p < .05 \). When asymptotic learning was compared by averaging the last six Category B trials in each condition, preference scores averaged 31.2% higher in the contrast condition, \( t(33) = 2.89, SE = 0.108, p < .01 \).

Although more instances of Category B were presented in the practice condition than in the contrast condition, subjects in the contrast condition showed greater learning of Category B. Presenting four instances of Category B during the pretraining block of the practice condition strongly interfered with the later learning of both categories in that condition. According to the category invention theory, this interference was due to inadequate learning of Category A defaults prior to encountering the first instance of Category B, which caused subjects to aggregate both types of instances into a single category.

In summary, the results of the present experiment were consistent with category invention and cannot be accommodated easily within a strictly autocorrelographic approach. The only qualification of this support for category invention derives from the temporary increase in the listing of Category A defaults that occurred after the first instance of Category B was presented in the contrast condition. There are several ways to interpret this slight readjustment of Category A norms at the start of the test block. In theory, the instance of Category B should have triggered the invention of a new category and thus have had no effect on Category A norms nor on preference scores for subsequent instances of Category A. One possibility is that the first instance of Category B triggered a new category as expected, but that the instance was assimilated both to this new category and to Category A. The new category would then provide a better match to subsequent instances of Category B than would Category A, so for these later instances only the new Category B would be evoked. Meanwhile, the Category A norms would gradually return to previous levels as subsequent instances of Category A were assimilated and overwhelmed the effects of the earlier Category B values.

A related possibility (L. W. Barsalou, personal communica-

### Experiment 3

This experiment was a modification of Experiment 2 designed to further investigate category invention in unsupervised learning. In particular, the present experiment investigated the influence of initially aggregating two contrast categories into a single class on subjects' ability to subsequently acquire accurate category-level discriminations.

All conditions of this experiment resembled the contrast condition of Experiment 2, except that the series of same-category instances in the pretraining block remained defined by a single instance from the contrasting category. In the contrast condition of Experiment 2, eight instances of Category A had been presented in succession prior to a mixed block of both Category A and Category B instances. Those eight instances were sufficient for most subjects to learn strong Category A defaults prior to encountering the first instance of Category B,

---

7 Such early contrasting of the two categories could not have been detected in Experiment 1 because only instances of Category B were presented during the second block of that experiment. This would have made it impossible to observe any temporary changes in Category A norms during that interval.
thus causing a new category to be created upon seeing the first instance of Category B. In the present experiment, rather than presenting all instances of Category A during the pretraining block, a single instance of Category A was presented on the first trial, followed by a series of instances of Category B (by convention, the category presented first in the training sequence is always referred to as Category A). The main independent variable in this experiment was the number of Category B instances that followed the first instance of Category A in the pretraining block; one group of subjects had 4 instances of Category B in this series, a second group had 8, and a third group had 12. After this pretraining block, a mixed block of both Category A and Category B instances, similar to that of Experiment 2, was presented for the next 13 trials.

The objective of presenting instances from two different categories on the first two trials was to cause subjects to aggregate the categories at the start of training. Because Category A was presented first, the aggregate norms should have initially been dominated by the values of that Category A instance. As subsequent instances of Category B were presented, however, the consistent features of that category should have competed with, and then dominated, the contrasting Category A values in the aggregate norms. If sufficient instances of Category B occurred in this series, these Category B values would be learned as defaults of the combined category, so that presenting a second instance of Category A would trigger a new category to accommodate it. The result of more instances of Category B, then, would be rapid learning of both Categories A and B during the subsequent mixed block.

By contrast, if insufficient Category B instances occurred prior to the test block, the probability of creating a new category should have been reduced. This reduction would result from the relatively high residual strengths of the Category A values in the aggregate norms, which would lessen the perceived contrast between those norms and the features of the first instance of Category A of the test block. If, as predicted, such subjects perceived little disparity and failed to segregate the second instance of Category A from the aggregated norms, that failure would be revealed in their attribute listings during the final mixed block, when they should show reduced learning of both categories.

Autocorrelation models predict a different pattern of results. Consistent with category invention, in such models one would expect that increasing the number of instances of Category B in pretraining should increase later Category B learning, simply because of increased practice. However, in autocorrelation theory one would expect that this manipulation would also decrease later Category A learning because of negative transfer or interference at the level of correlational associations or rules. Thus, the autocorrelation theory is inconsistent with improved learning of Category A because of the increased number of instances of Category B presented during pretraining.

**Procedure.** The procedures for this experiment were identical to those of the previous two experiments, except that the numbers of trials differed. Subjects were tested for a single half-hour session in groups of 8 to 10. They were given test booklets similar to those used in Experiments 1 and 2 and were allowed to complete the listing task at their own pace. The listing instructions were identical to those used in Experiments 1 and 2.

*Materials and design.* The stimuli in this experiment were the same pictorial insect stimuli used in Experiments 1 and 2. The stimulus set was partitioned into categories on the basis of perfectly correlated values on five binary attributes, as in Experiment 2. The remaining three attributes varied independently over two values, different for the two categories. The design shown in Table 2 for Experiment 2 held true for Experiment 3.

The main difference between Experiment 3 and Experiment 2 was the order in which training instances from the two categories were presented. The first instance was always different from the second; following the conventions of previous experiments, we refer to the instance presented first as belonging to Category A. The following n instances were from Category B; the number of instances in this series was the independent variable in this experiment. These first n + 1 instances (one Category A instance plus n Category B instances) were referred to as the pretraining block. This pretraining block was followed by a mixed test block of seven Category A and six Category B instances presented in random order (with the constraint that no more than two instances of the same category could occur in a row).

Each of the 16 possible instances from the training set was presented at least once, and instances were selected for a second or third presentation such that each value of the variable attributes appeared equally often. As in Experiment 2, two different stimulus sets were prepared such that assignment of default or variable status to a given attribute was balanced across the group of subjects; this balancing is depicted in Table 2. For both stimulus sets, booklets were constructed such that one category of insects played the role of the first-presented Category A for some subjects, whereas other subjects received booklets in which the other set of insects played the role of Category A. Crossing these two balancing factors (the stimulus set used and the order in which categories were presented) with the three levels of the n variable (the number of Category B instances in the pretraining series) yielded a total of 12 groups. Three subjects were randomly assigned to each group, for a total of 36 subjects in this experiment.

**Results and Discussion**

Preference scores for the three conditions of this experiment are shown in Figure 5. The main prediction of category invention tested in the present experiment was that increasing the number of instances of Category B in pretraining would increase learning of both categories in the following mixed block. This was expected because increasing the number of Category B instances should increase the relative strength of Category B values in the aggregated norms while decreasing the residual strength of Category A values from the first trial. This, in turn, should increase the probability of creating a new category when the next instance of Category A is encountered because these Category A values should appear relatively surprising with respect to these aggregated norms. Once the categories were disaggregated by this triggering, default learning could occur rapidly for each.

The pattern of results shown in Figure 5 lends support to these expectations. Preference scores for Category A (numbered A2 through A8 in Figure 5) increased significantly during the test block for conditions n = 12, (11) = 3.68, SE =

---

**Method**

**Subjects.** The subjects were 36 undergraduate students of Stanford University participating in partial fulfillment of an introductory psychology course requirement.
Turning to the test block, one can see that Category B learning was again higher in the $n = 12$ condition and lower in the other two conditions. Preference scores for the $n = 12$ condition exceeded those of the $n = 8$ condition by 27%, a significant difference, $t(22) = 2.14, SE = 0.126, p < .05$. In addition, preference scores were 22% higher in the $n = 12$ condition than in the $n = 4$ condition, a marginally significant effect, $t(22) = 1.73, SE = 0.126, p < .10$. No significant difference was obtained between the $n = 4$ and $n = 8$ conditions, $t(22) = 0.41, SE = 0.123$. When the $n = 4$ and $n = 8$ conditions were pooled into a single condition, preference scores in this condition were significantly less than those in the $n = 12$ condition, $t(34) = 2.28, SE = 0.107, p < .05$.

Although some learning may have occurred in all three groups, the stronger learning observed in the $n = 12$ condition favors the category invention theory over a pure autocorrelation theory. The present results, therefore, reinforce and expand on the results of Experiment 2 by demonstrating further patterns of transfer that appear incompatible with strict autocorrelation and that appear to require category invention. However, category invention does not predict higher learning in the $n = 4$ condition than in the $n = 8$ condition, which appeared to have occurred here; rather, we had expected a monotonic increase in learning as $n$ was increased from 4 to 12. The most plausible interpretation of these results is that no real differences existed between the $n = 4$ and $n = 8$ conditions, only between these two conditions and the $n = 12$ condition. Although the $n = 8$ condition appeared to show slightly less learning in some comparisons than the $n = 4$ condition, these comparisons were not statistically significant. Moreover, it appears likely from these data that the baseline learning ability of subjects assigned to the $n = 4$ condition was higher than that of subjects in the other two conditions. When we compared an interval of pretraining trials shared by all three groups (the second- to fourth-presented instances of Category B), we found that learning was significantly higher in the $n = 4$ condition than in the $n = 8$, $t(22) = 3.40, SE = 0.084, p < .01$ and in $n = 12$ conditions, $t(22) = 2.70, SE = 0.789, p < .01$.

Comparisons of Category B learning showed an ordering of conditions similar to those of Category A. Within the pretraining block, learning appeared greater in the $n = 12$ condition than in the $n = 8$ and $n = 4$ conditions, but not greater in the $n = 8$ than the $n = 4$ condition. Preference scores on the last pretraining trial were marginally greater in the $n = 12$ condition than in the $n = 8$ condition, $t(22) = 2.06, SE = 0.076, p < .10$, nonsignificantly greater in the $n = 12$ condition than in the $n = 4$ condition, $t(22) = 1.47, SE = 0.094, p > .10$, and not significantly different between the $n = 8$ and $n = 4$ conditions, $t(22) = 0.75, SE = 0.111, p > .20$. The results appeared slightly stronger when only default listings from the final pretraining trial were compared. Default listing was significantly less in the $n = 12$ than the $n = 8$ condition, $t(22) = 2.27, SE = 0.098, p < .05$ and in the $n = 4$ condition, $t(22) = 2.24, SE = 0.062, p < .05$, but there was no significant difference between the $n = 8$ and $n = 4$ conditions, $t(22) = 0.81, SE = 0.103, p > .25$.

Figure 5. Preference scores for the three conditions of Experiment 3. Pretraining trials are shown in their original order, whereas the test block trials are separated by category.
A Role for Autocorrelation?

whenever possible but that learning may occur by autocorrelation. How does the theory explain learning in mixed conditions? Experiment 1), is yet to be explained. If categories must be separated early in training for category invention to occur, then how does the theory explain learning in mixed conditions?

One possibility is that separate categories are created whenever possible but that learning may occur by autocorrelation. Our results do not imply that such autocorrelation is never a factor in unsupervised learning, only that this process alone cannot account for the sequence effects observed here. One possible hybrid theory would incorporate both category invention and autocorrelation; according to such a theory, subjects would normally accumulate some information about interfeature correlations as they processed successive training instances in a discrimination task. This would enable them to learn correlational patterns eventually, even without explicit category invention. Such a correlation learning process might be relatively slow because a large matrix of interfeature correlations would have to be learned in order for a category to be acquired. Consistent with this prediction, learning in the mixed conditions of these experiments was slower than that observed in the blocked conditions.

Although the present results do not eliminate the possibility of explicit autocorrelation in the mixed conditions, the results can also be explained without such autocorrelation. In principle, strict separation of categories in the training sequence should not be required for category invention to occur. For example, it might be assumed that learners in our experiments may invent a new category with some probability, \( P \), whenever an instance is presented that is from a different category than the stimulus presented on the trial before (e.g., when an instance of Category B is presented on a trial following one or more instances of Category A). The value of \( P \) would depend, in part, on the strength of the Category A default values in the norms for the single category that had been applied to all instances up to that point in training. In a blocked sequence, \( P \) would be high when the first instance of Category B occurred because subjects would have learned strong Category A defaults prior to encountering this instance. In a mixed sequence, \( P \) would be lower because both Category A and Category B defaults would be encoded as routine values in a set of aggregated norms, and neither would cause a radical mismatch with these norms nor a high probability of inventing a new category when they were presented.

However, category invention could still occur in a mixed sequence so long as the value of \( P \) was not too low. To illustrate, imagine that there was a 10% chance that the learner would create a new category whenever an instance of Category A occurred after an instance of Category B, or vice versa. Assuming that instances from Categories A and B were presented in alternation, the probability of creating a category on or before the \( n \)th alternation is \( 1 - (1 - P)^n \), where, for \( P = .1 \), reaches 53% by the sixth alternation. However, category invention would occur at different times for different subjects. Some subjects might discriminate between categories virtually from the start of training, others might do so later in the sequence, and a few might fail to do so by the end of a given training session. The data from such a process, averaged over a group of subjects, would show much the same pattern of apparently gradual learning predicted by the autocorrelation theory.

In summary, these experiments cannot discriminate between pure category invention and a hybrid theory that includes both category invention and autocorrelation. Although the present data provide evidence for the existence of a
category invention process, they cannot be interpreted as evidence for the nonexistence of autocorrelation.

**Generality of the Results**

One objection to generalizing from these results to unsupervised learning in the real world is that the stimulus variation in these experiments was rather artificial and stereotyped compared with the rich, complex variation typical of real-world domains. This objection applies to almost all laboratory research on category learning, which typically uses artificial stimuli generated from combinations of as few as two or three attributes. The purpose of the present experiments was to evaluate the attribute-listing task as an index of unsupervised learning in a relatively simple situation and to use it to make elementary discriminations among models of learning in that situation. Demonstrating that a process such as category invention occurs under artificial conditions constitutes a perfectly valid proof of the existence of that process, although it leaves the issue of boundary conditions unexplored.

In principle, the basic attribute-listing method could be used with many types of stimuli, including stimuli more complex and naturalistic than those used in the present experiments. However, complex stimuli should not change the basic pattern of results (i.e., a shift away from listing predictable aspects of the stimuli with an increasing focus on unpredictable information as categories are learned).

Another sense in which the present stimuli appeared artificial was in the fact that the default values of each category occurred with 100% reliability, that is, attributes were perfectly correlated. These experiments did not attempt to demonstrate unsupervised learning of categories with probabilistic defaults, which may limit the generality of the present results. However, the attribute-listing procedure should be generalizable to learning problems in which category defaults are somewhat unreliable, assuming that people can learn such categories without feedback. It is clear from subjects’ performance in the present task that many of the fuzzy categories used in standard supervised learning experiments, in which diagnostic features are often highly unreliable (e.g., Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Homa, 1984; Medin & Schaffer, 1978), might be very difficult for subjects to learn without explicit feedback. This difficulty in learning would simply reflect the greater difficulty of the unsupervised learning task itself, which requires subjects to generate their own categories and internal feedback. The investigation of such issues should provide interesting topics for future research and will allow useful comparisons between supervised and unsupervised learning.

**References**


Received July 9, 1992
Revision received May 24, 1993
Accepted June 11, 1993