AD-A261 658

DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

DTIC ELECTE APR 08 1993 S B

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S) AFOSR-TR- 93 0073
6a. NAME OF PERFORMING ORGANIZATION Department of Psychology 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION same as 8a.
6c. ADDRESS (City, State, and ZIP Code) Stanford University Stanford, CA 94305	7b. ADDRESS (City, State, and ZIP Code) same as 8c.
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Air Force Office of Scientific Research 8b. OFFICE SYMBOL (If applicable) NL	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER AFOSR-91-0144
8c. ADDRESS (City, State, and ZIP Code) Building 410 Bolling AFB DC 20332-6448	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 61102F; PROJECT NO. 2313; TASK NO. A4; WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)
Spontaneous Discovery and Use of Categorical Structure

12. PERSONAL AUTHOR(S)
John P. Clapper, Gordon H. Bower

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Annual Technical	FROM 01/15/92 TO 01/14/93	1993, February 15	31

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	unsupervised learning, category invention, attribute, value, autocorrelation, feature, default, variable
05	10		

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

These experiments investigated unsupervised category learning using tasks in which subjects attempted to memorize the features of training instances from two contrasting categories. On each trial, subjects studied a verbal feature list (training instance) for 24 seconds, after which they were given multiple choice recognition tests to evaluate their memory for each list item. The amount of time spent looking at each feature during the study phase, and the accuracy of recognition during the test phase, provided two separate indices of unsupervised learning on each trial. The main independent variable in these experiments was the specific sequence in which instances from the two categories were presented. The effects of these sequence manipulations on learning provided strong evidence for the use of an explicit, non-incremental, "category invention" process to capture the consistent structure of the stimulus domain. The present experiments also showed the selective encoding process and enhanced memory for instances predicted by standard, schema-based, theories of learning.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL John F. Tangney, Ph.D.	22b. TELEPHONE (Include Area Code) (202) 767-5021
	22c. OFFICE SYMBOL AFOSR/NL

DD Form 1473, JUN 86 Previous editions are obsolete. SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

93 3 5 037

12 FEB 1993

93-04816

31 pg

Abstract

This research investigates the unsupervised learning of categories, how such learning is affected by the sequencing of training instances, and how it alters and improves the encoding and retention of information about particular instances. Two general approaches to unsupervised learning are described, one based on learning explicit associations among correlated features (autocorrelation) and the other based on creating separate categories without explicit learning of correlational rules or associations (category invention). A "study time" procedure was used as an index of learning in these experiments; category learning is revealed in this task by subjects' preference to study features that differentiate among instances within a category while neglecting predictable features shared by all category members. These experiments obtained strong evidence for the use of a non-incremental category invention process in unsupervised learning. In addition, such learning improved subjects' ability to remember both expected and unexpected information about individual instances.

I. Research Objectives and Summary of Progress

This project aims to investigate the learning of categories in unsupervised tasks, in which no external tutor is present to provide subjects with pre-defined categories and informative feedback. This has involved several subgoals. First, we have developed new task paradigms and dependent measures for investigating unsupervised learning; this was necessary due to a lack existing measures of such learning. Second, these tasks have been employed to help discriminate between two rival theoretical frameworks describing how categorical structure could be learned and represented in unsupervised domains. One approach, which we refer to as "autocorrelation", relies on learning direct associations between correlated features of category members, without partitioning the stimulus set into explicit categories. The other approach, referred to as "category invention", is based on dividing the input stimuli into explicit categories and then computing summary norms within each category. A third objective of this research was to describe how category knowledge, once acquired, alters and improves the evaluation, encoding, and retrieval of information about individual category members.

In the first year of funding, we focused mainly on a task referred to as "attribute listing", in which subjects were presented with a series of training instances (pictures of fictitious insects), and asked to list the distinguishing properties of each instance. These lists were then analyzed over trials to reveal subjects' induction of generic norms about the experimental categories. An article describing several of these experiments is currently in press with the *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

We have developed a second task paradigm for investigating unsupervised learning, which we refer to as the "study time" task. This task consists of presenting subjects with a series of verbal stimulti (lists of features possessed by fictitious tree species) and instructing them to study and attempt to memorize the features in each list. Following a 24 second study period, a series of multiple choice recognition questions is presented to evaluate subjects' memory for the features of the preceding instance. Subjects are only allowed to look at one feature at a time during the study period, and a computer program records how long they spent studying each one. The program also records their accuracy for each item on the multiple-choice tests. As subjects learn the consistent, default, features of each category, they spend less time studying these predictable defaults and more focusing on the unpredictable variables. The decrease in study times to defaults and the corresponding increase to variables provides an index of unsupervised learning over trials that closely corresponds to that provided by the attribute listing procedure mentioned above. Interestingly, the recognition accuracy data provides a similar record of subjects' learning; accuracy of verifying both default and variable features increases as subjects learn the consistent features of each category.

One set of experiments was primarily concerned with discriminating between the autocorrelation vs. category invention approaches to unsupervised learning. These experiments manipulated the particular sequence in which training instances from two different categories were presented, and compared the effects of these manipulations to those predicted by the competing theories. These experiments were similar to some of the attribute listing studies briefly referred to above, and the data from these new experiments (both study times and recognition accuracy data) were highly consistent with those earlier results. That is, they provided strong evidence for the use of category invention in unsupervised learning, and showed sequence effects that could not be accommodated by autocorrelation. Some of these experiments are described more fully in the detailed report which follows.

A possible criticism of both the attribute listing and the study time experiments mentioned so far is that they all employed categories in which default features occurred with 100 percent reliability, whereas many real-world categories are characterized by fuzzy boundaries and unreliable defaults (e.g.,

Wittgenstein, 1953; Rosch, 1975, 1977). In a second set of study time experiments, we have begun to extend this procedure to investigate unsupervised learning of categories with probabilistic defaults. In one experiment, subjects were presented with instances of a single category, characterized by a set of default attribute values that each occurred in 90 percent of the instances, but were replaced by "exceptional" values in the other 10 percent. After several trials subjects showed much greater study times to surprising, exceptional, values than to predictable defaults. They also showed a slight "dishabituation" effect in which an attribute with a default value received longer slightly longer study times on a trial following the occurrence of an exceptional value on that attribute. These results imply that the study time procedure may be used to investigate unsupervised learning of categories with probabilistic defaults, which could greatly extend the generality of this research.

Two additional experiments were conducted to check whether the sequence manipulations investigated in earlier attribute listing and study time experiments would have the same effects when categories were characterized by probabilistic, rather than deterministic, defaults. The results of these experiments were generally consistent with those earlier results, providing further evidence for a non-incremental category invention process in unsupervised learning. Work is presently continuing on these issues.

A third area of research has involved using the attribute listing and study time tasks to study the acquisition of multi-layer conceptual hierarchies in unsupervised domains. As people acquire expertize within a given domain, they learn rich hierarchies of interrelated categories and subcategories at multiple levels of specificity. Such hierarchies may provide a foundation for inferences based on property inheritance, as well as efficient memory organization and fact retrieval. There have been few demonstrations of learning of multi-level categories or even reliable methods for observing such learning, especially within unsupervised learning tasks.

A first study time experiment attempting to demonstrate unsupervised learning of a simple two-layer hierarchy has produced encouraging results. The stimuli in this experiment were divisible into two general categories (A vs. B); category A could then be further divided into two more specific subcategories, which we referred to as A1 and A2. We found that subjects were able to learn default expectations at both superordinate and subordinate levels of generality, and that this learning considerably improved their memory for the features of individual instances.

Experiments during the 1993 funding year will be aimed at several issues. First, we wish to further investigate and clarify the conditions required for category invention, as well as other learning processes such as autocorrelation, particularly as they apply to learning categories with probabilistic defaults. Second, we plan to extend our initial work on multi-layer conceptual hierarchies, in particular investigating the progressive learning and elaboration of more specific (subordinate) categories within a domain and the organization of the resulting database in memory. And third, we plan to extend the study time task to obtain reaction time as well as accuracy data from the recognition-memory tests. These reaction times should be useful for investigating how information about categories and instances is organized in memory. In particular, we plan to follow up earlier results described in Clapper & Bower (1991) suggesting an explicit segregation between category and instance information in memory; such segregation would have important consequences for information storage and fact retrieval.

II. Publications

1. Clapper, J. P. & Bower, G. H. (1991). Learning and applying category knowledge in unsupervised domains. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation, Vol. 27*, Academic Press, New York, pp. 65- 108.

2. Clapper, J. P. & Bower, G. H. (in press). Category invention in unsupervised learning. *Journal of Experimental Psychology: Learning, Memory and Cognition.*

3. Clapper, J. P. & Bower, G. H. "Instance and category learning in unsupervised tasks". Manuscript to be submitted shortly.

III. Participating Personnel

1. Gordon H. Bower, PI
2. John P. Clapper, Research Associate
3. Katherine Longueville, Research Assistant

IV. Detailed Report of Study Time Experiments

A detailed description of study time experiments designed to differentiate between autocorrelation and category invention follows.

Instance and Category Learning in Unsupervised Tasks

The ability to learn and use categories is fundamental to human intelligence. Categories may be acquired under two general classes of training conditions, referred to as *supervised* and *unsupervised* learning. In a typical supervised learning experiment categories are defined in advance by the experimenter, who also provides relevant feedback (reinforcement) so that subjects can gradually learn to match these categories to the correct class of training instances. By contrast, in unsupervised learning tasks subjects are not given predefined categories or feedback from an external tutor. Rather, subjects must discover categories for themselves as they examine a series of training instances, basing such categories on any patterns or regularities observed among these stimuli.

A rich research tradition has evolved in the study of supervised learning (see, e.g., Goodnow, Bruner, & Austin, 1956; Millward, 1971; Smith & Medin, 1981), but there have been comparatively few empirical studies of unsupervised learning. One reason for this paucity of research may have been a lack of reliable measures of category learning within such tasks. For example, accuracy in choosing among a set of predefined categories, the primary measure used in studies of supervised learning, is by definition inapplicable to unsupervised learning.

Clapper and Bower (1991, 1993) developed and tested an index of unsupervised learning, referred to as "attribute listing". In the present article, we introduce a second method for investigating unsupervised learning; this procedure generates two distinct indices of learning on each training trial. This new method employs the same basic strategy or approach as the attribute listing task, and is based on similar assumptions. Below, we briefly review the earlier attribute listing studies, their underlying assumptions, and how attribute listing was used to provide discriminating tests between two competing theoretical approaches to unsupervised learning. We then describe the new task, showing how it may provide converging evidence concerning the rival theoretical approaches, and in addition provide information about how category induction alters and economizes the processing of individual instances.

Measures of Unsupervised Learning

One empirical strategy, described in Clapper and Bower (1993), is to study unsupervised category learning within *instance discrimination* tasks, by using the priority or weighting given to different features of the presented stimuli as an indirect index of category learning. This approach depends on two assumptions: (1) categories are defined in terms of correlated (consistently co-occurring) properties within a stimulus domain; and (2) correlated properties are mutually redundant for distinguishing among individual instances within a domain, and so they should receive a lower weighting or attentional priority than uncorrelated properties.

Regarding the first assumption, we begin by adopting a conventional vocabulary describing training instances in terms of abstract dimensions or *attributes*, each of which can assume a number of concrete *values* (Clapper & Bower, 1991, 1993). For example, people differ in the attribute of hair color, with blond, brown, red, and black being possible values of this attribute. A specific value of an attribute possessed by a given instance is also referred to as a *feature* of that instance. In principle, attributes may be either additive (with two values, present and absent) or substitutive (with any number of alternative values, such as the different hair colors listed above; see, e.g., Tversky, 1977). Attributes may also be discrete or continuous (e.g., ordered dimensions such as height or weight). In this article, only the discrete, substitutive case will be considered, although the methods described should also be applicable to other cases.

Given a stimulus domain described in terms of a particular set of attributes, categories may be defined within this domain in terms of correlations among the values of these attributes (see Figure 1). Such correlational structure (Garner, 1974) provides an inductive basis for partitioning a domain into separate categories, each corresponding to a particular set of correlated features. Importantly, it also provides the learner with predictive power -- given that one or two correlated values are observed, the presence of the others can be readily inferred. [1] To the extent that a subject discovers and learns such correlational patterns without feedback or other assistance from an external tutor, we consider that unsupervised learning has occurred.

Insert Figure 1 about here

Regarding the second assumption listed above, we argue that the learning of correlation-based categories can be studied using tasks in which subjects' instructed goal is to learn to discriminate among (identify) the individual training instances, i.e., in which category learning is not presented to subjects as an explicit goal of learning (Clapper & Bower, 1993). In such instance discrimination tasks, the objective is to learn unique responses to each individual instance, which in turn depends on learning how that instance differs from all other presented stimuli. Each feature (attribute value) of an instance would be evaluated in terms of its informativeness or utility for making such discriminations.

Within a particular stimulus set, there are two factors which, in principle, would determine an attribute value's discriminative informativeness: (1) the *probability* that an instance possessing that value is the target instance and not a lure, i.e., the proportion of lures eliminated by possessing that attribute value rather than an alternative value, and (2) the *redundancy* of the discriminations provided by the present feature with those provided by other features. If two attribute values are perfectly correlated within a domain, then they distinguish the target instance from identical sets of lures, and discrimination would not be improved by knowning both values rather than only one.

Insert Figure 2 about here

A rational or ideal subject in a such a task should allocate attention (cognitive capacity) among the features of an instance on the basis of their discriminative informativeness. Specifically, features that provide little discriminative information should receive a low attentional priority (weighting). The attributes within the stimulus domain illustrated in Figure 2 are equated in terms of their baseline probability of occurrence, but differ in their degree of redundancy. Mutually redundant, correlated, values should be regarded as less informative than the uncorrelated values, and should therefore receive a lower priority. If subjects did in fact pay less attention or otherwise assign a lower priority to these correlated values, this would be evidence that they had internalized the correlational patterns. Hence, an observable index of feature weighting could provide an indirect index of learning correlational patterns in unsupervised tasks.

In Clapper and Bower (1991, 1993), attribute listing was used as an index of feature weighting. Specifically, subjects were presented with a series of training instances (pictures of fictitious insects) and asked to write down the features that would be required to distinguish each one from prior instances they had seen. They were told not to list features that would be uninformative for such discriminations, even if the omitted features were highly prominent or noticeable. Subjects in this task preferred to list uncorrelated features over correlated features; this preference evolved gradually over trials as subjects had the opportunity to discover and learn the correlational patterns within the stimulus sets. This preference

was interpreted as a quantitative index of learning, and could be plotted over trials to display acquisition functions for each category.

Theoretical Approaches to Unsupervised Learning

We distinguish two general approaches to learning and representing in memory the types of correlational patterns depicted in Figures 1 and 2; these approaches follow directly from our definition of categories in terms of correlational patterns.

First, the correlations may be represented directly, as a set of correlational rules or within a correlational matrix. This approach is illustrated by some of the connectionist models of J. A. Anderson (Anderson, 1977; Anderson, Siverstein, Ritz & Jones, 1977) and McClelland and Rumelhart (1985; Rumelhart, McClelland & the PDP Research Group, 1986). It is also instantiated in rule-based systems such as those of Billman and Heit (1988) and Davis (1985). We will refer to this as the *autocorrelation* approach (Clapper & Bower, 1993). By keeping a record of the correlations between all possible pairs of attribute values, a learner could capture the correlational structure of stimulus sets like those in Figures 1 and 2 without actually partitioning the domains into explicit categories. Any information that would be provided by such a classification would already be implicit in an exhaustive correlational record; in fact, explicit categorization would actually lose or obscure certain correlational information by averaging over individual correlations to arrive at a single number for each attribute value (that value's probability of occurrence within the category).

The second approach is to capture the correlational patterns by partitioning the stimulus set into separate categories, as shown in Figure 1. General norms or expectations about each category are then stored in separate data structures, such as prototypes or schemas. There are many theories that assume that people represent category norms within such structures (e.g., Posner & Keele, 1968; Reed, 1972; Minsky, 1975; Rumelhart & Ortony, 1977; Schank & Abelson, 1977; Schank, 1982; Anderson, 1991), although most were not specifically intended to handle unsupervised learning. The schema or mental model of each category is assumed to contain generalizations about the range of expected values for each attribute. When a particular value is present in all or most of the instances within a category, subjects learn to expect that value to occur in future instances; we refer to such highly expected values as the *default* values of a category. By contrast, uncorrelated values that occur infrequently or probabilistically within a category will be referred to as *variables*.

By sorting stimuli containing different correlational patterns into different categories, and then computing averages or frequency distributions within these categories, it is possible to capture much of the same information contained in a direct correlational record. We refer to such theories as the "category invention" approach to unsupervised learning. Whereas autocorrelation models require only a single learning process (for updating correlational rules or associations), category invention requires two distinct processes, one for partitioning the conceptual space into separate categories and the other for computing norms across instances within each category (Michalski & Stepp, 1983).

It is probably unrealistic to assume, as do many statistical clustering models (see, e.g., Michalski & Stepp, 1983; Fried & Holyoak, 1984), that human learners can scan an entire set of stimuli at once and then compute an optimal classification scheme based on this overall analysis. It is more realistic to portray people as examining a set of training instances one at time and updating their conceptual knowledge in response to each. Given this sequential learning assumption, the major practical issue faced by the category invention approach is deciding when, and on what basis, to create new categories during training.

When the goal is to learn category summaries or schemas, and sequential learning is assumed, then a learner must use the match or mismatch of each stimulus to existing categories to decide when to invent a new category (e.g., Schank, 1982; Holland et al., 1986; Anderson, 1992). We assume that subjects create a new category at the start of an experiment to describe the first training instance. Further training instances are then assimilated to this reference category until an instance is encountered that mismatches the category in excess of some internal criterion. When this occurs, the subject creates a new category to describe the anomalous instance. If further instances similar to this initial "triggering" instance are later encountered, they will also be assigned to the new category. Separating the norms for different categories in this way allows new patterns to be learned without discarding or distorting knowledge of old patterns.

In Clapper and Bower (1993), the attribute listing task was used to provide discriminating tests of the autocorrelation versus category invention theories of unsupervised learning, described above. These tests depended on the vulnerability of the category invention process to initial distortions or errors in learning, depending on the particular sequence in which training instances are presented. The data showed that learning was much better if one category was learned thoroughly prior to encountering any instances of the other category. Under such conditions, the mismatch between the well-learned norms of the first category and the contrasting features of the second category was highlighted, and subjects readily learned to separate them. (This was reflected in a rapidly-evolving preference for noting uncorrelated variables over correlated defaults in the listing task). By contrast, learning was greatly reduced when instances of both categories were presented together, in a mixed input sequence, from the start of training. In this case, the contrast between the two categories was apparently much less salient, and it appeared that many subjects simply lumped all the stimuli together into a single, overgeneralized category. Because this single category averaged together instances containing different correlational patterns, such correlational information would have been lost in the aggregated norms. Subjects in such mixed sequence conditions showed much less preference for listing variables over defaults than did subjects who learned the categories separately.

Perceived contrast does not affect learning within the autocorrelational approach, since such models simply increment correlational strengths without imposing any classification scheme upon the stimulus domain. In other words, autocorrelation is a strictly data-driven ("bottom up"), inductive, learning method, without the potential for distortions or errors implicit in the inherently theory-driven ("top down") process of partitioning a domain into separate categories. Autocorrelation models do not necessarily expect superior learning when categories are separated in the training sequence, compared to situations in which they are presented in mixed alternation.

Autocorrelation models could be constructed in which different correlational patterns interfered with each other's learning; this would be consistent with much research on associative interference in paired associate learning and sentence memory tasks (see, e.g., Postman, 1971; Anderson, 1983). Such interference could explain why a category might be learned better if presented alone than if presented in a mixed sequence with instances of a different category. However, it does not explain several results reported in Clapper and Bower (1993) which are readily explained by the category invention approach. For example, interference effects should occur in both blocked and mixed sequences, according to this interference hypothesis. In fact, certain connectionist autocorrelators predict much greater interference in blocked than in mixed sequences (McCloskey, 1989; Ratcliff, 1990). By contrast, evidence of significant interference was obtained only in the mixed conditions of these experiments. Other apparent violations of incremental correlation learning were also observed; for example, under certain circumstances learning of a category could be improved simply by reducing the number of instances presented from that category, a result difficult to accommodate within a strict autocorrelational framework. Overall, the results of these experiments were strongly supportive of category invention, and could not easily be rationalized in terms of simple autocorrelation.

A Performance-Based Measure of Unsupervised Learning

A primary goal of the present research was to extend the results reported in Clapper and Bower (1993) to a new task in which the indices of learning were based on actual performance and capacity limitations, rather than on subjects' preference for including one type of attribute rather than another in a free listing task. However, this task is based on the same principles that underlied attribute listing, e.g.[, that subjects would assign greater weight to uncorrelated than correlated attribute values when trying to distinguish among individual training instances.

Subjects were presented with training instances composed of several attributes, some of which had correlated values (defining two contrasting categories), and some of which did not. In the attribute listing studies, the training stimuli were pictures of fictitious insects; in the present experiments, they were lists of verbal features supposedly possessed by different species of trees. For example, a given tree species might be described as having dark grey bark, a high commercial value, fast growth, and so on. Subjects were required to study these feature lists for a fixed study interval; during this time, the display was set up so that the person could only look at one feature at a time. After the study period, subjects were tested on their ability to recognize which features had occurred in the previous instance, i.e., for each attribute such as "bark color", the subjects would have to decide which of several alternative values (e.g., dark grey, deep brown, mossy green, or light tan) occurred in the last instance.

These lists were presented on a microcomputer screen, which allowed two types of data to be collected: (1) the time spent looking at each attribute value during the study period, and (2) the accuracy of verifying each value during the testing phase. Interestingly, both types of data provide information about category learning similar to that provided by attribute listing. Thus, we expected that subjects who learned the categories within a given stimulus set would spend more time studying variables than defaults, because the variables were more distinguishing of each instance and because these features could not be inferred based on category norms or correlational rules. This preference for studying variables over defaults would be given the same interpretation as the corresponding preference for listing variables over defaults in the attribute listing task, i.e., as indications of learning categories or correlational patterns.

A second index of learning was provided by the recognition-memory data in the present experiments. Subjects who learn categories should show improved memory for defaults, since they would be able to retrieve these features from generic norms when they were needed for the memory tests. Interestingly, subjects should also show improved memory for variables, compared to a control condition in which all attributes of the stimuli are uncorrelated. This improvement should occur as a result of the preference, predicted above, for increasing the portion of the study period spent looking at (rehearsing) variables at the expense of defaults. This extra study time should improve subjects' memory for variables, without affecting verification accuracy for defaults. Thus, default learning should produce both the direct benefit of improved memory for defaults, and the indirect benefit of better memory for variables (see Clapper & Bower, 1991).

In sum, the present instance-memory task was designed to provide two measures of unsupervised learning on each trial, both consistent with the earlier attribute listing measure. In addition to providing similar information about the time course of category learning, the present task provides additional information about how category learning affects the processing of and memory for individual training instances. This is important because category and instance learning do not appear to be totally independent processes. Clapper and Bower (1991) argued that the changed processing of instances that results from category learning (i.e., the shift of attention away from predictable defaults and toward

unpredictable or surprising properties of the instance) could facilitate the learning of further categories within a domain. This might occur both as a result of improved instance memory (better "raw data" obviously permit more accurate and reliable generalizations), and because subjects would be more likely to discover subtle, non-obvious features and patterns within a stimulus domain once they shifted their attentional resources away from the more obvious defaults. We argued that these attentional shifts were an important factor underlying the heightened episodic memory (e.g., deGroot, 1965, 1966; Chase & Simon, 1973), and progressive elaboration of default hierarchies (see Holland, Holyoak, Nisbett, & Thagard, 1986) shown by domain experts.

Overview

The goals of the following experiments were two-fold.

First, we hoped to provide evidence for the basic validity and usefullness of the instance memory procedure as a method of investigating unsupervised learning. To do this, we conducted two experiments similar to attribute listing studies described in Clapper and Bower (1993). If the results of these experiments were consistent with those of the earlier attribute listing studies, this would provide evidence for the reliability of both tasks and the basic stability of the underlying processes they attempt to investigate.

The generality of our methods and theoretical conclusions would be further bolstered by the fact that the present experiments differed from the earlier attribute listing studies in several ways. For instance, the present studies used verbal stimuli with a larger number of attribute dimensions than were employed in the pictorial attribute listing stimuli. It is important to include both verbal and pictorial stimuli in research on unsupervised learning because previous research indicates that verbal stimuli may be remembered (Pavio, 1971; Kosslyn & Pomerantz, 1977) and compared (Gati & Tversky, 1984) differently than pictorial stimuli, which could also mean that they are categorized somewhat differently.

Our second objective was to provide further evidence relevant to discriminating between the autocorrelation versus category invention approaches, described above. The earlier attribute listing studies provided strong support for the category listing position, which we hoped to replicate in the present experiments. To that end, the main independent variable in the present experiments was the particular sequencing of training instances. If the present sequencing manipulations replicate those of Clapper and Bower (1993), this replication would strengthen the case for a non-incremental, contrast based process of category invention.

Experiment 1

The main goals of this first experiment were to evaluate the instance memory task as an index of unsupervised learning, and to provide evidence to discriminate between the category invention versus autocorrelation theories. There were three conditions in this experiment. In two of these the stimulus set was partitioned into contrasting categories (A versus B) based on correlations among the values of nine attributes, while the remaining three attributes varied independently. These are referred to as correlated conditions. The same stimuli were presented in both of the correlated conditions; the only difference between them was the particular order in which training instances occurred. In the ***Blocked*** condition, a block of twelve A-instances was followed a second block of twelve B-instances. Following these two "pure" blocks was a mixed test block consisting of four instances from each category, presented in random order. In the Mixed condition, the same first twenty-four instances were presented as in the Blocked condition, but these instances were presented in random order rather than being separated by

category. The same test block was used as in the Blocked condition.

The third condition was a control group. The stimuli were equated with those of the correlated conditions in the number of values associated with each attribute, but there were no correlated values, and hence no categories, in this group. Thus, this condition served as a baseline for evaluating any learning observed in the other two groups.

The two correlated conditions provided a test of the category invention versus autocorrelation theories. As noted above, category invention expects better learning when instances are blocked by category, because this allows subjects to learn strong expectations about Category A prior to encountering the first instance of Category B. Category invention predicts that subjects should have difficulty separating categories in the Mixed condition, and that they would be likely to aggregate both types of instances into a single overgeneralized category containing no strong default expectations. If this occurred, then subjects should show a greater preference for studying variables over defaults, as well as better memory for both defaults and variables, in the Blocked condition,.

The autocorrelation framework can accommodate reduced learning in a Mixed sequence (compared to a condition in which categories are learned alone) by including assumptions about interference among correlational rules or associations. However, if such an interference process reduced learning in the Mixed condition, it should also influence the pattern of results from the Blocked condition. First, prior learning of Category A should interfere with later learning of Category B in the Blocked condition, analogous to the negative transfer (or proactive interference) commonly observed in paired-associate learning tasks (e.g., Postman, 1971). Second, correlation learning during the Category B block should produce retroactive interference on earlier learning of A correlations, causing a reduction in A-learning during the final test block.

Method

Subjects

The subjects were 43 undergraduate students of San Jose State University participating in partial fulfillment of their Introductory Psychology course requirement.

Procedure

Subjects were tested in groups of 10 to 15 for a single one-hour session. Each subject was seated in front of an individual microcomputer terminal, which administered all aspects of the experiment. After subjects read the instructions presented on the computer screen and signed a form indicating their informed consent to participate, the main portion of the experiment began.

Each trial consisted of two phases, the study phase and the test phase. At the beginning of the study phase, a list display was presented in the middle of the CRT screen. At the top of the list was the name of a fictitious tree species (these were arbitrarily selected Latin names from a plant identification guide), below which appeared a list of twelve verbal feature descriptors. At the start of the trial, each descriptor was masked by a row of X's (see Figure 3a). Starting from a random position in the list, subjects studied the descriptors by pressing a designated "line forward" or "line backward" key to examine each list item. This allowed subjects to examine the features in any order they wished, and to spend as much time as they wished on any particular item within the constraints of the prespecified study

period (24 seconds). The computer recorded the total amount of time spent looking at each attribute.

Insert Figure 3 about here

Each list item was a verbal description of a specific value of a particular stimulus attribute. For example, the attribute "color of bark" had several alternative values, such as "dark grey" and "mossy green". The attributes were presented in the same serial order on each trial, although different values of a particular attribute could occur on successive trials.

After a study interval of 24 seconds, the list disappeared and the test phase of the trial began. During this test phase, subjects were tested on their memory for all twelve of the attribute values of the preceding instance. The test items were presented one at a time in a multiple-choice format (see Figure 3b). The name of the most recent instance appeared at the top of the multiple-choice display, with four alternative answers below. These alternatives were always different values of the same attribute, e.g., four different habitat preferences or growth rates. Subjects decided which of these values occurred in the last-studied instance and typed in the number corresponding to that choice on their computer keyboard. Following this response, the computer displayed either a "correct" or an "incorrect" prompt under the test display, which remained on the screen. If the response was incorrect, the correct choice was indicated by an arrow in the display (see Figure 3c). A designated key was then pressed to show the next test question.

After they had answered all twelve test questions about a given instance, subjects received summary feedback for the trial. The percentage of items answered correctly on that trial was displayed, and below this the cumulative percentage correct averaged over all test trials completed up to that point. If the trial score was higher than the cumulative score, the message "Good job! You beat your overall score!" appeared on the screen; if not, the message "Try to beat your overall score next trial" was displayed. If the subject answered all the test questions correctly on a given trial, the message "Good job! Your score was perfect!" was displayed.

The twelve attributes were tested in a different random order on each trial, and the order in which values were listed in the multiple-choice display was also randomized separately on each trial. The experiment consisted of a total of 32 such study-test trials. Following this, a written debriefing was shown which informed subjects about the purpose and methods of the experiment.

Materials and Design

As noted, the training instances were verbal descriptions of fictitious trees, presented in a list format. The instances were characterized in terms of twelve substitutive attributes, each of which had four possible values, defining a a possible stimulus set of 4^{12} distinct instances. For nine of these twelve attributes, only two of the four possible values were presented in the training instances, although all four values appeared as responses in the multiple choice tests.

Subjects were randomly assigned to three different conditions. In the two correlated conditions the values of the nine two-valued attributes were perfectly correlated across different training instances. The instances could be partitioned into two distinct subsets or categories based on these correlated values. These can be denoted by letting serial positions in a numerical sequence correspond to particular attributes, while the numbers appearing in those positions indicate specific values of each attribute. Within this notation, the categories can be described as Category A = 111111111xxx and Category B = 222222222xxx, where the x's indicate uncorrelated attributes that vary independently through all four

values across different instances of a category. As noted above, the correlated values characteristic of a given category are referred to as *defaults*, while the values of the non-correlated attributes are called *variables*.

The two correlated conditions differed in the order in which instances were presented. In the *Blocked* condition, the first twelve instances were all members of Category-A and the second twelve instances were members of Category-B. The remaining eight trials consisted of four A-instances and four B-instances presented in a randomly intermixed sequence. The *Mixed* condition differed from the Blocked condition only in the order in which the first twenty-four instances were presented. In this condition, these instances were presented in a randomly ordered sequence rather than being blocked by category. The randomization procedure was so constrained that no more than three instances from the same category appeared in a row. The final eight trials were identical to those of the Blocked condition.

The third condition in this experiment was referred to as the uncorrelated or *Control* condition. In this condition, all the attributes of the training instances varied independently. As in the correlated groups, nine of the twelve attributes varied through only two values in the training instances, while the remaining three attributes varied through four values. Due to the lack of correlations among attribute values in this condition, there was no structural basis for partitioning the stimuli into separate categories. A total of $2^9 \times 4^3 = 32{,}768$ distinct instances are possible in this condition, compared to $4^3 \times 2 = 128$ possible instances in the correlated conditions.

The final eight instances presented in the Control condition were identical to those of the two correlated conditions. That is, these instances contained correlated values, unlike the preceding twenty four instances. This final block of correlated instances will be referred to as the *test block* in all three groups.

Balancing

The stimuli for all the subjects in a given condition were generated by the testing program from the same input file, which contained coded specifications for generating the instances presented on each trial. Stimuli generated from these codes were presented in the same order in which they occurred in the file, i.e., in the same order for all subjects in a given condition. The correspondence between serial positions in the codes and the order in which an attribute was listed in the training instances was randomized for each subject. These random assignments were undertaken to balance out any idiosyncratic effects of particular attributes, values, or combinations of values on the experimental data.

Results and Discussion

The two dependent variables recorded on each trial of this experiment were (1) study times for default and variable attributes during the study phase, and (2) recognition accuracy for defaults and variables during the test phase. Since the total duration of the study period was a constant 24 seconds,[2] any increase in study times (STs) to variables would be reflected in a corresponding decrease in default STs. Therefore, in this article the ST results will be described in terms of the *difference* in study times between variables and defaults on a given trial, i.e., ST(variables minus defaults). Following Clapper & Bower (1993), we will refer to these differences as *preference scores*, since they reflect subjects' preference for attending to variables rather than defaults. The data for this experiment are shown in Figure 4.

Insert Figure 4 about here

Beginning with the *Blocked* condition, the mean ST for defaults was 1.78 seconds and that for variables was 2.91 sec, for an average difference of 1.13 sec. This difference was highly significant according to a within-subjects t-test, $t(14) = 4.27$, $p < .001$. Examining the difference scores plotted over trials in Figure 4a, it is apparent that the bias in favor of studying variables increased throughout the A-category block, from .18 sec on the first trial to 2.01 sec on the twelfth and final trial of this block. The within-subjects test for a linear trend during this block was statistically significant, $t(14) = 2.86$, $p < .02$. This learning did not appear to reach assymptote by the twelfth trial, and more learning might have been observed if additional A-instances had been presented prior to the Category B block.

Learning seemed to occur somewhat more rapidly during the Category B block, and reached assymptote by about the 6th B-instance. Default STs exceeded variable STs on the first B-trial by 0.125 seconds; the decrease in difference scores from 2.01 on the final trial of the A-block to -0.125 sec on the first B-trial was highly significant, $t(14) = 4.01$, $p < .01$. The increased learning over the first six B-instances was significant at the .01 level, $t(14) = 4.04$, but no significant change occurred over the next six B-instances, $t(14) = -0.37$, $p > .50$. The trend computed over all twelve trials of the B-block was also significant, $t(14) = 4.48$, $p < .01$.

The bias in favor of attending to variables decreased somewhat when the first A-instance was presented during the mixed test block, compared to the average of the preceding six B-instances ($t(14) = 3.71$, $p < .01$), but it is clear from Figure 3a that preference scores remained positive throughout the test block. This effect was highly significant averaged over the eight test trials, $t(14) = 3.05$, $p < .01$. This is an important result because it indicates that the learning effects observed earlier in the training sequence were not due merely to localized habituation to "runs" of repeated default values, but rather to the acquisition of stable norms for the two categories.

The autocorrelation-plus-interference hypothesis, described earlier, predicts that learning of a second category in a blocked sequence should produce strong retroactive interference on memory for the first. Such interference implies that preference scores during the test block should be lower in instances of Category A than in B-instances. However, excluding the first A-instance, there was no significant difference in preference between A- versus B-instances during the test block, $t(14) = 0.04$, $p > .50$. The slightly lower preference scores for instances of both categories during this block, compared to the eight preceding B trials ($t(14) = 2.62$, $p < .05$), were probably due to the need to sample enough of the default features to confidently categorize the instance on each trial of the test block. During the earlier blocks, category membership was constant over long series of trials, and thus subjects may have spent less time checking the categorization of each instance during these trials.

Turning to the *Mixed* condition, no significant difference was observed between variable and default STs (means of 2.04 and 2.07 sec, respectively, $t(14) = 0.60$, $p > .50$). The preference scores showed no apparent trends over the thirty two trials of the experiment; any variation appears merely due to random fluctuations from trial to trial. The data for the uncorrelated *Control* condition were similar to those of the Mixed condition. Variable STs averaged only about .06 sec greater than default STs, a non-significant effect ($t(12) = 0.669$). There were no significant learning trends in this condition.

In addition to the foregoing within-groups analyses, several between-groups analyses were undertaken to directly compare the different conditions. The average preference score of 1.14 seconds observed in the Blocked condition was significantly greater than the 0.06 second effect observed in the Control condition, $t(26) = 3.63$, $p < .01$. The same comparison was were also statistically significant when averaged over only the eight-trial test block, $t(26) = 2.81$, $p < .01$. Preference scores in the Blocked

condition also exceeded those in the Mixed condition overall ($t(28) = 4.29$, $p < .001$) and during the test block ($t(28) = 2.78$, $p < .01$). No comparison between the Mixed and Control conditions approached significance.

The pattern of study time results was strongly replicated by the recognition memory data (Figure 4b). In the Blocked condition, recognition improved for both defaults and variables over the first several trials of both the A- and B-blocks. Averaged across defaults and variables, overall accuracy increased from 0.66 on the first A-instance to an assymptote of about 0.92 on the ninth trial. Accuracy dropped to 0.71 on the first B-trial; the difference between this trial and the preceding A-trial was significant at the .001 level, $t(14) = 6.47$. A similar pattern of increasing accuracy was observed over the succeeding B-instances.

The increasing linear trend in accuracy was significant over the first six instances of both categories ($t(14) = 3.94$, $p < .01$ for Category A, $t(14) = 4.71$, $p < .001$ for Category B). By contrast, there was no significant trend over the last six instances of either category (for Category A, $t(14) = 1.12$; for Category B, $t(14) = -1.32$). A slight decrease occurred during the first few trials of the mixed test block, and overall memory performance during this block differed somewhat from assymptotic performance during the preceding Category B block (computed by averaging the last six trials of that block and comparing this mean to the average of the eight test trials; $t(14) = 3.29$, $p < .01$). However, when the first A-instance was excluded there was no overall difference in memory between the two categories during this test block, $t(14) = 1.07$, $p > .15$. Thus, there was little evidence for strong retroactive interference of Category B on memory for defaults of Category A.

While the overall pattern of results over trials was similar for defaults and variables, memory for defaults was greater overall (0.93 vs 0.83, $t(14) = 5.45$, $p < .001$). This advantage could have been due to (1) subjects' ability to retrieve correlated default values from their category norms, while the values of variable attributes had to be recorded from scratch for each instance, or (2) the greater ease of guessing the correct value of attributes that had only two values presented during the study phase, compared to those that had four presented values.

By contrast, there were no clear trends in the memory data from the Mixed condition. Overall, defaults were recognized with an average accuracy of 0.65 and variables with an accuracy of 0.60; this difference was significant at the .01 level ($t(14) = 3.01$). Since there is no other evidence of default learning in the data, it seems likely that this difference was due to the greater ease of guessing the correct value of two-valued as compared to four valued attributes, rather than to subjects having learned the correlations among the two-valued attributes.

In the Control condition, memory was at about the same level as in the mixed condition (0.620 versus 0.625, respectively), and showed no clear changes over trials. Recognition was about eight percent more accurate for two- than for four-valued attributes, comparable to the corresponding difference in the Mixed condition. This difference was statistically significant at the .01 level ($t(14) = 3.83$).

Directly comparing memory accuracy from the Blocked vs. Control conditions, we found that accuracy in the Blocked condition was significantly greater than that of the Control condition ($t(26) = 7.07$, $p < .001$). When the recognition data was separated into defaults vs. variables, accuracy was greater for both types of attributes in the Blocked condition. This improvement averaged 27 percent for defaults ($t(26) = 8.53$, $p < .001$) and 24 percent for variables ($t(26) = 5.44$, $p < .001$). The amount of improvement for defaults did not significantly exceed that for variables, $t(26) = 1.07$, $p > .10$.

The fact the category learning (in the Blocked condition) increased memory for both defaults and variables indicates that such learning facilitates encoding of both predictable and unpredictable features of instances. This replicates earlier results showing that category knowledge improves memory for both default and non-default properties of instances (Clapper & Bower, 1991), and provides support for the encoding assumptions of standard schema theories and their variants. Such theories assume that learners focus on those aspects of an instance that are surprising or unpredictable with respect to norms stored in the category schema, while ignoring or backgrounding expected defaults (see, e.g., Bower, Black, & Turner, 1979; Graesser, Woll, Kowalski, & Smith, 1980). This was what was observed in the study time data from the present experiment, and the recognition data provided further verification.

The overall pattern of memory data from the Mixed condition was very close to that of the Control condition, as would be expected from the ST data indicating that no category learning occurred in the Mixed condition. None of the comparisons between Mixed and Control group data approached statistical significance in this experiment.

To summarize, the pattern of results from both study times and verification accuracies show much better learning in the Blocked condition than in the other two groups, and this finding lends support to the category invention approach. There was no evidence for proactive interference due to learning Category A upon subsequent learning of Category B in the Blocked condition; in fact, assymptotic learning was reached at least as quickly in the second category as in the first. This lack of interference contradicts a prediction of autocorrelation, i.e., if interference occurs between categories in a mixed sequence, then it should also affect learning in a blocked sequence. The autocorrelation-plus-interference hypothesis also expects that learning of Category A during the test block should have been reduced by retractive interference from Category B. However, after the temporary surprise of seeing the first A-instance, subjects showed no difference in learning of the two categories during the test block. The present results are difficult to accommodate within a strictly autocorrelational framework, and imply that people in unsupervised learning tasks accommodate unfamiliar stimuli by inventing new categories.

Experiment 2

This experiment aimed to provide further evidence for category invention. Subjects were randomly assigned to two conditions. In the first, referred to as the *Contrast* condition, sixteen instances of Category A were presented prior to a mixed block of twelve A-instances and twelve B-instances. We expected that subjects in this group would learn strong defaults for Category A during the first, or pretraining, block, and that the contrast between these well-learned defaults and the features of the first B-instance would cause a new category to be invented when that instance was encountered at the beginning of the second, or test, block. Due to this partitioning, the defaults of Category B should be learned quickly and without interference from Category A in this group.

The *Practice* condition of this experiment was essentially a replication of the Mixed condition from Experiment 1. Here, eight A-instances and eight B-instances were presented in random order during pretraining, after which the same mixed test block of twenty four A- and B-instances was presented as in the Contrast condition. In this case, category invention models expect that subjects would have difficulty perceiving the contrast between the two categories, and be likely to assimilate both types of instances to a single set of aggregated norms. The result would be greatly reduced learning, compared to the Contrast condition.

Autocorrelation predicts a different pattern of results, particularly with regard to the learning of Category B. Eight instances of Category B were presented during pretraining in the Practice condition, whereas no B-instances occurred in the pretraining block of the Contrast condition; the same number was

presented to both groups during the test block. Due to this larger number of instances, learning of Category B should be superior in the Practice condition. This prediction can be derived not only on the basis of greater practice of B correlations, but also from a consideration of expected interference (transfer) effects. A larger number of A-instances are presented in the first block of the Contrast condition than in the Practice condition; this should result in greater interference upon subsequent B-learning, and, again, better learning of Category B in the Practice condition.

Category invention predicts that transfer in this experiment should be *positive* from Category A to Category B, i.e., B-learning should be improved by increasing the number of A-instances in the pretraining block from eight in the Practice condition to sixteen in the Contrast condition. At the same time, transfer from Category B to Category A should be *negative*, i.e., replacing eight of the A-instances presented in the Contrast condition with eight B-instances, as in the Practice condition, should decrease later learning of Category A. These seemingly contradictory predictions make little sense within the framework of simple autocorrelation, but are easily rationalized in terms of category invention.

Method

Subjects

The subjects were 31 students of San Jose State University participating in partial fulfillment of their Introductory Psychology course requirement.

Procedure

The experimental procedure was identical in most respects to that of Experiment 1. Subjects were tested in groups of 10 to 15 for a single session lasting approximately one hour. Each subject was individually seated at his or her own computer terminal in a single large testing room. The entire experiment, consisting of 40 trials plus instructions and debriefing, was administered by computer.

Materials

The tree description were designed according to the same general specifications used in Experiment 1. Each instance (individual species) was described in terms of twelve attributes, and the stimulus set was partitioned into two categories based on correlations among the values of nine of these twelve attributes. These categories can be denoted as Category A = 111111111xxx and Category B = 222222222xxx, where each serial position represents a particular attribute, 1 and 2 are the default values of Categories A and B, respectively, and the x's indicate attributes that vary independently through all four possible values. The assignment of particular attributes to the default or variable condition was performed randomly for each subject.

Design

Subjects were randomly assigned to two conditions, which differed only in the sequencing of the training instances. In the Contrast condition, instances of Category A were presented for the first sixteen trials, referred to as the pretraining block. Following this pretraining, a mixed test block was presented in which twelve instances of each category were presented in a random order (these sequencings were re-

randomized for each subject). In the Practice condition, the pretraining block consisted of a mixed block of eight A-instances and eight B-instances presented together in a random order. The same mixed test block was used as in the Contrast condition.

In both conditions, instances were so constructed that all four values of each variable attribute occurred an equal number of times within each category; within this constraint, values of these attributes were assigned randomly. The same stimulus set was presented to all subjects in a given condition, but the order of specific instances within the pretraining and test blocks was re-randomized for each subject.

Results and Discussion

The same type of data was collected in this experiment as in Experiment 1. This data is displayed in Figure 5.

Insert Figure 5 about here

We begin with analyses of the Contrast condition. The ST data showed strong evidence of learning in this condition. Overall, variables were studied 1.33 seconds longer than defaults; this preference was significant at the .001 level, $t(16) = 4.11$. Recall that only instances of Category A were presented during the pretraining block in this condition. During this time, preference scores increased from -0.16 on the first trial to 2.08 sec on the sixteenth trial. A within-subjects contrast computed over this interval showed a significant linear trend ($t(16) = 2.72$, $p < .02$). Thus, strong learning of A-norms appears to have occurred during pretraining.

Following the pretraining block (i.e., after the first B-instance had been presented), preference scores appeared to decrease for the first few A-instances of the test block. However, this decrease did not attain conventional levels of statistical reliability. For example, when comparing the last three trials of pretraining to the first three A-trials of the test block, no significant difference was observed (2.00 sec vs 1.65 sec; $t(16) = 1.22$, $p > .10$). Comparisons between various other intervals of trials in this region of the training sequence also failed to show a significant change in ST preference scores. Linear contrast analyses reveals no increasing or decreasing trend in the subsequent A-trials during the test block ($t(16) = 0.70$, $p > .40$).

Preference scores did decrease significantly on the trial when the first B-instance was presented, compared to the preceding A-trial (2.08 sec vs -0.19 sec, $t(16) = 3.90$, $p < .01$). This means that subjects regarded the new defaults of the B-category as highly informative on that trial, and allocated them equal attention to the variables. The linear trend over the twelve B-instances in the test block was highly significant ($t(16) = 4.31$, $p < .001$), implying strong learning of the B-norms during this block.

Overall, the ST data for the Blocked condition show strong learning of Category A during the pretraining block, no significant reduction of this A-learning during the test block, and strong B-learning during the test block.

The Practice condition was essentially a replication of the Mixed condition from Experiment 1, and produced similarly little evidence of significant learning. Overall, four-valued attributes were studied slightly longer than two-valued attributes in this condition, but this difference did not approach statistical significance. The preference scores averaged 0.23 seconds overall ($t(13) = 1.69$, $p > .10$), 0.33 seconds for Category A ($t(13) = 1.69$, $p > .10$), and 0.13 sec for Category B ($t(13) = 1.28$, $p > .10$). Thus, there was

no statistical evidence of learning in the ST data from this condition.

In summary, strong evidence for category learning was obtained in the Contrast condition but not in the Practice condition. This difference in learning was further supported by direct statistical comparisons between the two groups. The mean ST preference score of 1.33 seconds in the Contrast condition was significantly greater than the corresponding 0.23 second preference in the Practice condition ($t(29) = 2.89$, $p < .01$). When this comparison was restricted to the test block (which was identical in both conditions), the effect remained highly significant ($t(29) = 3.01$, $p < .01$). The differences between the Contrast and Practice conditions were also significant when the two categories were analyzed separately ($t(29) = 2.99$, $p < .01$ for Category A and $t(29) = 2.94$, $p < .01$ for Category B).

The memory data from the Contrast condition showed evidence of category learning similar to that of the ST analyses (Figure 5b). Defaults were recognized with a mean accuracy of 94.3 percent, compared to 83.7 percent for variables ($t(16) = 5.76$, $p < .001$). Accuracy changed over trials with a pattern similar to that of the ST data from this condition. When default and variable means were averaged, a linear contrast over the first eight trials of the pretraining block showed a highly significant increase in subjects' memory accuracy, from 48.5 percent on the first trial to 88 percent on the eignth trial ($t(16) = 8.35$, $p < .001$). Following this initial increase, memory for A-instances remained fairly stable thereafter. Accuracy decreased sharply on the first B-trial, compared to the preceding A-trial ($t(16) = 6.87$, $p < .001$). Following this, accuracy increased significantly over the first eight B-trials, from 68 to about 93 percent ($t(16) = 3.86$, $p < .01$). This pattern of gradually improving memory for both categories provides a converging measure of learning that is highly consistent with the ST measure described above.

Turning to the Mixed condition, recognition accuracy was significantly greater for defaults (71.8 percent) than for variables (63.9 percent), $t(13) = 3.19$, $p < .01$. Accuracy increased significantly over the first four trials ($t(13) = 2.31$, $p < .05$), and remained approximately stable thereafter. Since the ST data shows no evidence of learning in this condition, the greater accuracy in verifying defaults compared to variables was probably due to the greater ease of guessing the correct values of the defaults, as discussed for Experiment 1.

The conclusion that significant learning occurred in the Contrast condition but not the Practice condition was further supported by direct comparisons of recognition accuracy between the two groups. Accuracy was greater in the Contrast condition both for defaults ($t(29) = 8.04$, $p < .001$) and for variables ($t(29) = 4.66$, $p < .001$). Defaults were recognized 10.6% more accurately than variables in the Contrast condition, while the corresponding difference in the Practice condition was 6.9%. A direct comparison between showed no statistically significant different between these two effects ($t(29) = 1.31$, $p > .10$). The finding that memory for variables was improved about as much as memory for defaults is consistent with the fact that subjects in the Contrast condition spent more time attending to variables than defaults during the study period. Such an increase in study time to variables would be expected to result in improved verification.

The finding of better learning in the Contrast condition, especially of Category B, provides strong evidence in favor of category invention. Autocorrelation cannot accommodate the finding that decreasing the number of instances seen from a given category could increase learning of that category, as shown in the present experiment. A strictly autocorrelational approach also cannot account for the lack of interference between categories in the Contrast condition, compared to that which occurred in the Practice condition.

General Discussion

The present experiments, along with earlier attribute listing studies, provide strong evidence for the use of an explicit category invention process in unsupervised learning. In both of the present experiments, subjects were better able to distinguish between two categories when given the opportunity to thoroughly learn one category prior to being exposed to any instances of the other category. We interpret this result as due to a sort of "learned contrast" effect: When norms for one category are well-learned, it is easier to see the contrast between these norms and an instance from a different category. This, in turn, increases the likelihood that the person will create a separate category to describe this mismatching stimulus, rather than assimilating both types of instances to a single set of aggregated norms.

The autocorrelational approach was shown to be unable to accommodate the present results. In particular, it cannot explain how simple manipulations of the training sequence determined interference effects between the categories, creating strong interference in some conditions while completely eliminating it in others. It also cannot explain the finding in Experiment 2 that reducing the number of instances presented from a given category can greatly improve learning of that category.

The present data support the commmonsense observation that people invent new mental models in response to the failure or inadequacy of old ones. This is illustrated by scientific research, in which new theories are generally proposed in response to a mismatch between a pre-existing category (theory) and a particular instance or case (data) to which it is unsuccessfully applied (e.g., Popper, 1959). In this paper, we operationalized the "failure" of a model as the occurrence of improbable or surprising values instead of expected defaults -- analogous to seeing a pink, furry elephant when our norms for this category predict hairless, grey, skin, or to obtaining a set of measurements that contradict standard theory in a physics experiment. We assumed that people would not discard or throw away their previous norms when such exceptional cases are encountered, but that they would instead construct new norms to apply specifically to these cases. [3]

These results, together with the attribute listing results of Clapper and Bower (1993), provide strong evidence for the generality of the category invention process. Evidence for category invention has been obtained with three different measures (attribute listing, study time, and recognition accuracy) in two different tasks, and with two different stimulus types (pictures of objects versus verbal feature lists). The task demands also differed across the two types of experiments. The attribute listing task measured subjects' evaluation of different features according to the criterion of instance discrimination, but subjects were never asked to demonstrate actual memory performance in those experiments. By contrast, the indices employed in the present experiments were closely tied to actual discrimination performance. The recognition tests directly evaluated subjects' ability to remember how each instance differed from the others, and the study time index directly reflected how subjects allocated their attention while preparing for the recognition tests.

The present results are also strongly supportive of the general episodic processing assumptions of schema-type theories (see also Bower, Black, & Turner, 1979; Graesser, Woll, Kawalski, & Smith, 1980) and with the literature concerning episodic memory abilities of domain experts (e.g., deGroot, 1965, 1966; Chase and Simon, 1973). Schema theories usually assume that subjects encode an instance (e.g., descriptions of individuals based on personality stereotypes, or of routine activities based on internalized scripts) by referring to the generic schema in memory (by encoding some sort of "pointer" to that schema, e.g., Graesser et al.) and then encoding only those features of the instance that could not be predicted from the schema, i.e., that are inconsistent with schema defaults or that pertain to variable attributes for which no defaults have been learned. In the present experiments, this would imply that subjects should encode each tree description by encoding the category membership of the tree, and then selectively

recording those variable values not inferrable from this categorization. In other words, subjects should look at the default values only long enough to classify the instance, and should then spend the remainder of the study period focusing on variables. Our finding that subjects spent more time studying variables than defaults is generally consistent with these expectations of schema theory and its variants.

One advantage of such "schema-based encoding" of instances is that memory for each instance is improved; this was illustrated in the present experiments by the improved memory that occurred in the blocked conditions, in which subjects were best able to tell the categories apart. The improvement is due to the fact the subjects only need to learn the features of each instance that are not already stored as default expectations in their category norms. As subjects learn the default features of a category, and are better able to focus on variable features, memory for these variables increases, as does the accuracy of verifying defaults. This improved learning may provide a model explaining the much greater retention of detailed information within a given domain by people who are accomplished experts in that domain, compared to novices (deGroot, 1965, 1966; Chase & Simon, 1973). Experts have a finely elaborated system of categories and subcategories pertaining to their chosen domain, and these categories provide default assumptions against which particular situations can be matched and evaluated, increasing memory for both expected and unexpected information.

Another advantage of selectively ignoring default values once a stimulus has been categorized is that this frees attentional resources to attend to other, non-default, features of the instance. This, in turn, may facilitate the discovery of new regularities among these non-defaults, and might also lead to the discovery of new attributes (previously unnoticed dimensions of variation within a given stimulus domain). To illustrate, once having learned to separate oak trees from maple trees, a learner would be better able to attend to the more subtle properties that distinguish different types of oaks because they would no longer attend to features common to all oaks. In naturalistic learning, people often consider known categories as "background" and proceed to focus on finer distinctions between instances that might form a basis for learning more differentiated categories. Thus, the attentional backgrounding of expected features may play an important role in the development and elaboration of default hierarches by domain experts (e.g., Holland, Holyoak, Nisbett, & Thagard, 1986). The same backgrounding phenomenon would also facilitate feature discovery and improvements in so-called "perceptual learning" within a domain (see E. Gibson, 1963, 1969).

In addition to these theoretical issues, a major objective of this research was the development of the empirical methods or task paradigms themselves, because obtaining detailed records of empirical phenomena and regularities within a scientific domain necessarily precedes and supports substantive theorizing about that domain. The present methods should be applicable to the investigation of several issues related to unsupervised learning, e.g., how subjects determine criteria for inventing new categories in different situations, how this depends on factors such as prior learning, sequencing of training instances, stimulus structure, training conditions, mental set or task strategy and so on. The present memory tasks can also be applied to issues relating to use of category knowledge for learning instances, and how this would depend on the reliability of category defaults, the degree of match between an instance and category norms, and many other factors. These issues should provide productive topics for future research.

References

Anderson, J. A. (1977). Neural models with cognitive implications. In D. LaBerge, & S. J. Samuels (Eds.), *Basic processes in reading: Perception and comprehension*. Hillsdale, NJ: Erlbaum.

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review, 84*, 413-451.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior, 22*, 261-295.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*, 409-429.

Billman, D., & Heit, E. (1988). Observational learning from internal feedback: A simulation of an adaptive learning method. *Cognitive Science, 12*, 587-625.

Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology, 11*, 177-220.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley.

Chase, W. G., & Simon, H. A. (1973). The mind's eye in chess. In W. G. Chase (Ed.), *Visual information processing*. New York: Academic Press.

Clapper, J. P., & Bower, G. H. (1991). Learning and apply category knowledge in unsupervised domains. In G. H. Bower (Ed.), *The psychology of learning and motivation, Vol. 27*. New York: Academic Press.

Clapper, J. P., & Bower, G. H. (1993). Category invention in unsupervised learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, in press,* .

Davis, B. R. (1985). An associative hierarchical self-organizing system. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-15*, 570-579.

deGroot, A. D. (1965). *Thought and choice in chess*. Mouton: The Hague.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 234-257.

Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum.

Gati, I., & Tversky, A. (1984). Weighting common and distinctive features in perceptual and conceptual judgements. *Cognitive Psychology, 16*, 341-370.

Gibson, E. J. (1963). Perceptual learning. *Annual review of psychology, 14*, 29-56.

Gibson, E. J. (1969). *Principles of perceptual learning and development*. New York: Appleton.

Graesser, A. C., Woll, S. B., Kowalski, D. J., & Smith, D. A. (1980). Memory for typical and atypical actions in scripted activities. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 503-513.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning and discovery*. Cambridge, MA: MIT Press.

Kossylyn, S. M., & Pomerantz, J. R. (1977). Imagery, propositions, and the form of internal representations. *Cognitive Psychology, 9*, 52-76.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114*, 159-188.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation, Vol. 24*. New York: Academic Press.

Michalski, R. S., & Stepp, R. E. (1983). Learning from observation: Conceptual clustering. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach*. Palo Alto, CA: Tioga Publishing Company.

Millward, R. B. (1971). Theoretical and experimental approaches to human learning. In J. W. Kling, & L. A. Riggs (Eds.), *Experimental psychology, third edition* (pp. 905-1017). New York: Holt, Rinehart & Winston.

Minsky, M. (1975). A framework for representing knowledge. In P. H. Winston (Ed.), *The psychology of computer vision*. New York: McGraw Hill.

Pavio, A. (1971). *Imagery and verbal processes*. New York: Hold, Rinehart, and Wilson.

Popper, K. R. (1959). *The logic of scientific discovery*. New York: Harper & Row.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*, 353-363.

Postman, L. (1971). Transfer, interference, and forgetting. In J. W. Kling, & L. A. Riggs (Eds.), *Experimental psychology, Third Edition* (pp. 1019-1132). New York: Holt, Rinehart & Winston.

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review, 97*, 285-308.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3*, 382-407.

Rosch, E. (1975). Cognitive representation of semantic categories. *Journal of Experimental Psychology: General, 104*, 192-233.

Rosch, E. (1977). Human categorization. In N. Warren (Ed.), *Advances in cross cultural psychology, Volume 1*. Academic Press.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1* (pp. 45-77). Cambridge, Mass.: MIT Press.

Rumelhart, D. E., & Ortony, A. (1977). The representation of knowledge in memory. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), *Schooling and the aquisition of knowledge*.. Hillsdale, N. J.: Lawrence Erlbaum Associates.

Schank, R. C. (1982). *Dynamic memory*. Cambridge, UK: Cambridge University Press.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Hillsdale, N. J.: Lawrence Erlbaum Associates.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327-352.

Wittgenstein, L. (1953). *Philosophical investigations*. Oxford: Blackwell.

Footnotes

This research was supported by Air Force Office of Scientific Research Grant AFOSR-91-0144.

Requests for reprints should be sent to Gordon H. Bower, Department of Psychology, Stanford University, Stanford, CA 94305.

1. The correlation of features within a category need not be perfect for that category to have positive utility, since some predictive power is gained by recording any correlational patterns that recur with greater-than-chance reliability. This is a significant point, because the features of natural categories are generally considered to be probabilistic rather than deterministic (Wittgenstein, 1953; Rosch, 1975, 1977). This means that properties generally true of a category are subject to exceptions; for example, although the ability to fly is one of the most characteristic features of birds, there are a few species that lack this ability. However, flight occurs frequently enough in conjunction with the bundle of properties that define the category "birds" to remain a highly reliable generalization about the class as a whole.

2. The computer checked the elapsed study time whenever a subject looked at a different feature of a given instance. Thus, the list display could only disappear when the subject moved on to a different feature, but not while they continued to look at the same feature. Because of this, the total study time sometimes exceeded 24 seconds by a small amount. However, this slight discrepancy did not affect the pattern of results and will not be discussed further.

3. In this sense, the present learning may be a little different than that which occurs in scientific research, since in science new observations sometimes cause old theories to be completely discarded or reformulated. However, scientists, like other people, are quite conservative about discarding a favored theory that has worked well in the past, and will often modify or elaborate the theory to accommodate special cases, rather than giving the theory up. This conservative strategy is reasonable from the perspective of cognitive economy, since it allows old beliefs to be retained without the costly errors that would result from misapplying them, without the cognitive effort that would go into creating an entirely new theory.

Figure Captions

Figure 1. Sample stimulus sets illustrating how categories are defined in terms of correlated attribute values.

Figure 2. Sample stimulus sets illustrating how the current value of each attribute of the target instance distinguishes that instance from a particular set of lures. Note that the first five attributes, which are correlated defaults, all distinguish the target instance from exactly the same set of lures.

Figure 3. Computer display as it appeared during each phase of Experiments 1 and 2.

Figure 4. Study time and verification accuracy data from Experiment 1. In this figure, the function connecting the "O" points is from the Blocked condition, that connecting the "*" points is from the Mixed condition, and the "." points are from the Control condition. Trials are shown in their original order in this figure; the functions are disconnected to indicate where the A- and B-blocks are separated in the Blocked condition, and where the test block begins in all conditions.

Figure 5. Study time and verification accuracy data from Experiment 2. The "O" points are from the Contrast condition while the "*" points are from the Practice condition. Pretraining trials are shown in their original order, but test trials are separated by category in both conditions.

Stimulus Set #1

Attribute								Attribute							
1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
1	1	1	1	1	1	1	1	2	2	2	2	2	1	1	1
1	1	1	1	1	1	1	2	2	2	2	2	2	1	1	2
1	1	1	1	1	1	2	1	2	2	2	2	2	1	2	1
1	1	1	1	1	1	2	2	2	2	2	2	2	1	2	2
1	1	1	1	1	2	1	1	2	2	2	2	2	2	1	1
1	1	1	1	1	2	1	2	2	2	2	2	2	2	1	2
1	1	1	1	1	2	2	1	2	2	2	2	2	2	2	1
1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2

Category "A" : 1 1 1 1 1 x x x
Category "B" : 2 2 2 2 2 x x x

Stimulus Set #2

Attribute								Attribute							
1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
1	1	1	2	2	1	2	1	2	2	2	1	1	2	2	2
1	1	1	1	2	2	1	1	2	2	2	2	2	1	1	1
1	1	1	1	1	1	1	1	2	2	2	2	1	1	2	1
1	1	1	1	2	2	2	1	2	2	2	1	1	2	2	1
1	1	1	1	2	1	1	2	2	2	2	2	1	1	1	2
1	1	1	2	1	1	1	2	2	2	2	2	2	2	1	1
1	1	1	2	2	2	2	2	2	2	2	1	1	2	1	2
1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2

Category "A" : 1 1 1 x x x x x
Category "B" : 2 2 2 x x x x x

Figure 1.

Stimulus Set #3

Attribute								Attribute							
1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
1	1	1	1	1	1	1	1	3	2	3	2	4	4	3	4
1	1	1	1	1	1	1	2	4	3	3	4	2	4	3	3
1	1	1	1	1	1	2	1	2	4	4	4	2	3	4	3
1	1	1	1	1	1	2	2	4	3	2	2	4	3	3	3
1	1	1	1	1	2	1	1	3	2	4	3	4	4	4	4
1	1	1	1	1	2	1	2	2	4	3	3	3	4	4	3
1	1	1	1	1	2	2	1	4	3	2	3	3	3	3	4
1	1	1	1	1	2	2	2	3	3	4	2	3	3	4	4

Category "A" : 1 1 1 1 1 x x x
"Not - A" : y y y y y y y y

Stimulus Set #4

Attribute								Attribute							
1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
1	1	1	1	1	1	1	1	2	2	2	1	2	1	1	1
1	1	1	2	1	1	1	2	1	2	2	2	2	1	1	2
1	1	1	1	1	1	2	1	2	2	2	2	2	1	2	1
1	2	1	1	1	1	2	2	2	2	2	2	2	1	2	2
1	1	1	1	1	2	1	1	2	2	2	2	2	2	1	1
1	1	1	1	1	2	1	2	2	1	2	2	2	2	1	2
1	1	1	1	2	2	2	1	2	2	2	2	2	2	2	1
1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2

Category "A" : 1 1 1 1 1 x x x
Category "B" : 2 2 2 2 2 x x x

Figure 2.

	Attribute								
Instance	1	2	3	4	5	6	7	8	
	1	1	1	1	1	1	1	1	<-- Target Instance
2)	1	1	1	1	1	1	1	2	
3)	1	1	1	1	1	1	2	1	
4)	1	1	1	1	1	1	2	2	
5)	1	1	1	1	1	2	1	1	
6)	1	1	1	1	1	2	1	2	
7)	1	1	1	1	1	2	2	1	
8)	1	1	1	1	1	2	2	2	
9)	2	2	2	2	2	1	1	1	
10)	2	2	2	2	2	1	1	2	
11)	2	2	2	2	2	1	2	1	
12)	2	2	2	2	2	1	2	2	
13)	2	2	2	2	2	2	1	1	
14)	2	2	2	2	2	2	1	2	
15)	2	2	2	2	2	2	2	1	
16)	2	2	2	2	2	2	2	2	

Lures excluded by the current value of each attribute of the target instance :

1 :	9, 10, 11, 12, 13, 14, 15, 16
2 :	9, 10, 11, 12, 13, 14, 15, 16
3 :	9, 10, 11, 12, 13, 14, 15, 16
4 :	9, 10, 11, 12, 13, 14, 15, 16
5 :	9, 10, 11, 12, 13, 14, 15, 16
6 :	5, 6, 7, 8, 13, 14, 15, 16
7 :	3, 4, 7, 8, 11, 12, 15, 16
8 :	2, 4, 6, 8, 10, 12, 14, 16

a. Aralia

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
dark grey bark
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

Press INS or DEL to see next item

---

b. Aralia

1. deep brown bark
2. dark grey bark
3. mossy green bark
4. light tan bark

****
* Enter a number from 1 to 4 *
****

---

c. Aralia

1. deep brown bark
--> 2. dark grey bark
3. mossy green bark
4. light tan bark

INCORRECT

Arrow indicates correct choice

Press RETURN to go on

Figure 4.

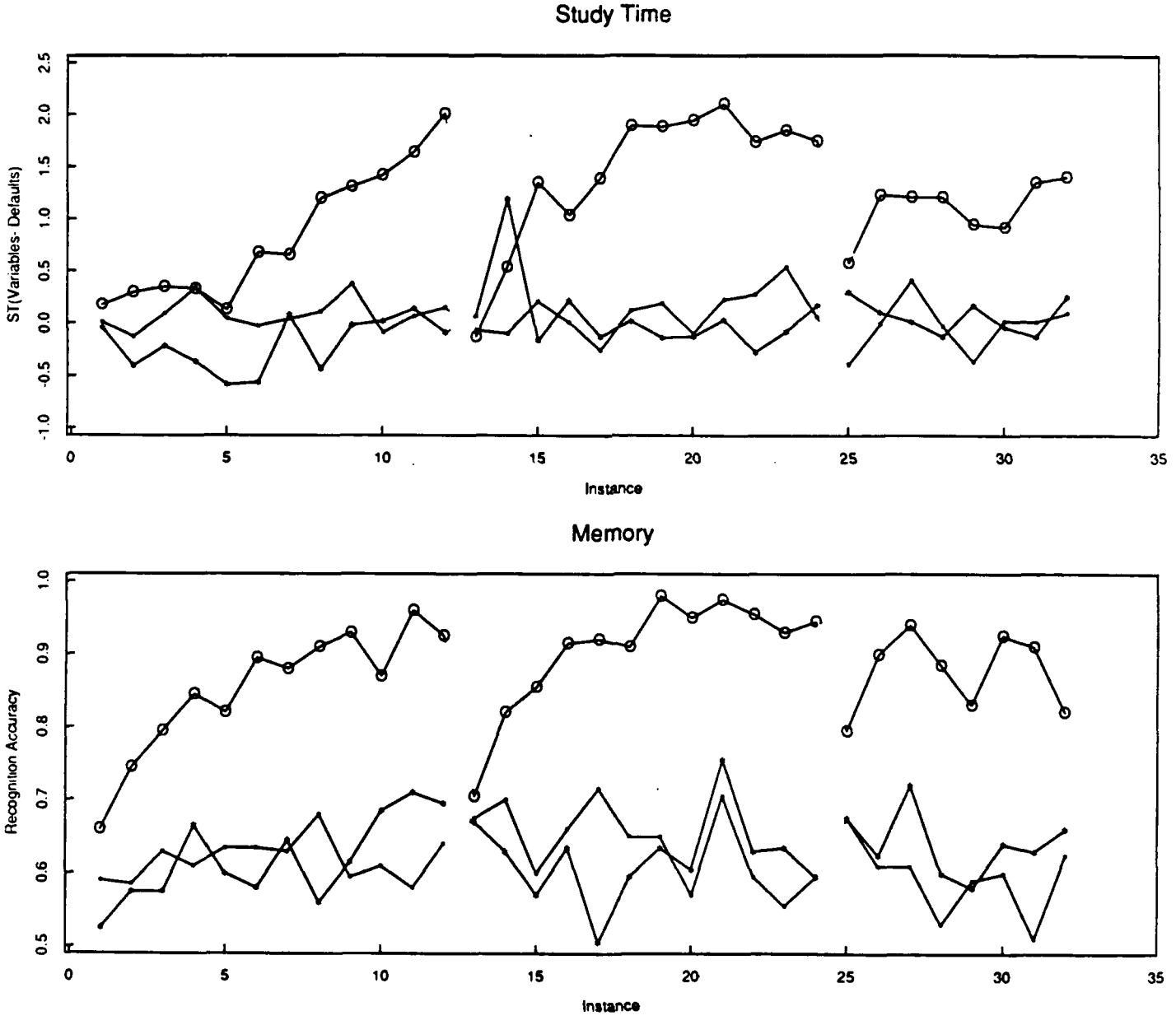
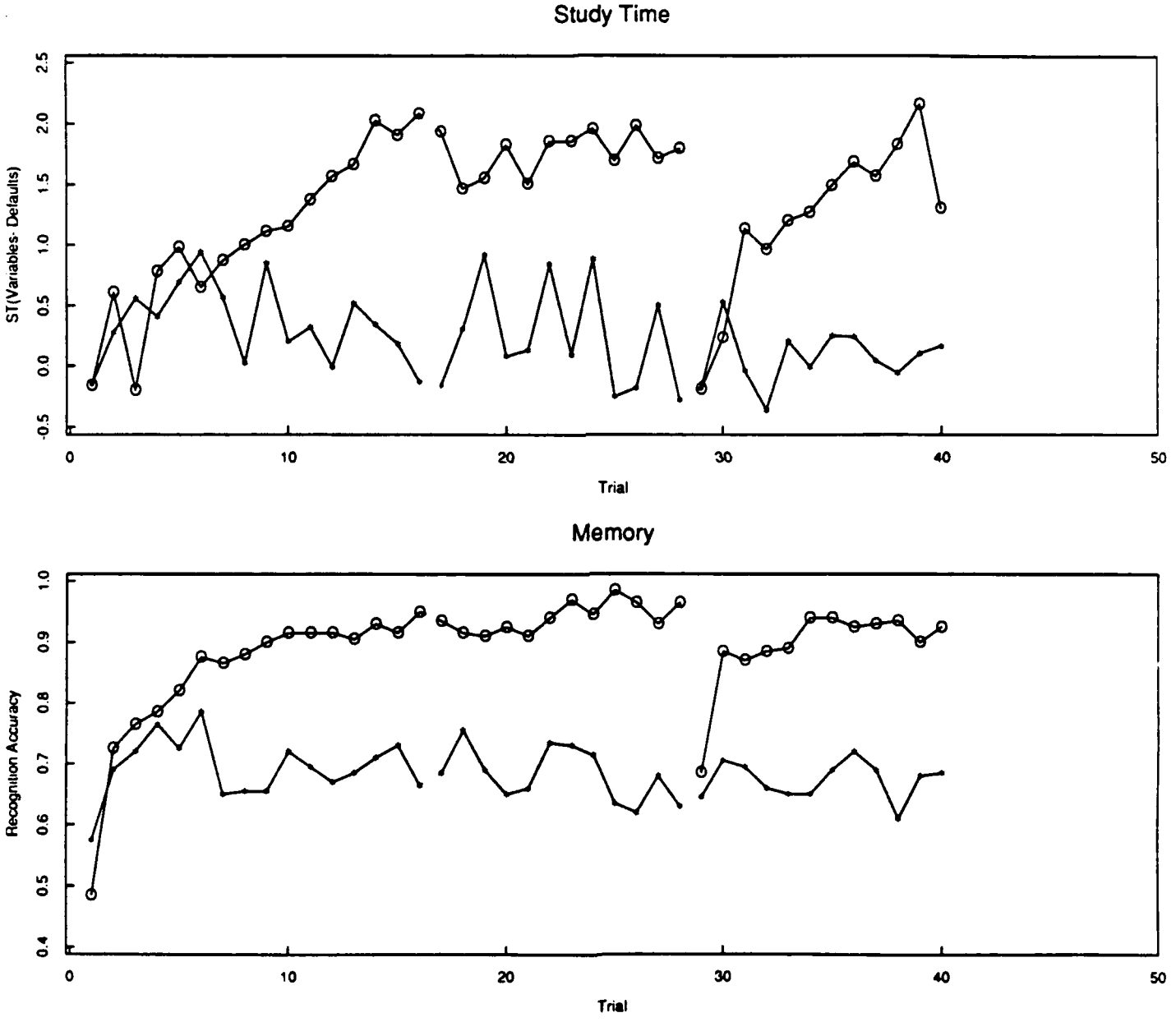

Study Time
ST(Variables- Defaults)
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
0
5
10
15
20
25
30
35
Instance
Memory
Recognition Accuracy
0.5
0.6
0.7
0.8
0.9
1.0
0
5
10
15
20
25
30
35
Instance

Figure 5.


Study Time
ST(Variables- Delaults)
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
0
10
20
30
40
50
Trial
Memory
Recognition Accuracy
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0
10
20
30
40
50
Trial