2

DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION
Unclassified

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3. DISTRIBUTION / AVAILABILITY OF REPORT
Approved for public release;
distribution unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)

5. MONITORING ORGANIZATION REPORT NUMBER(S)
AFOSR-TR- 92-0195

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.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
61102F	2313	A4	

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

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

13a. TYPE OF REPORT
Annual Technical

13b. TIME COVERED
FROM 01/15/91 TO 01/14/92

14. DATE OF REPORT *(Year, Month, Day)*
1992, February 15

15. PAGE COUNT
40

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP
05	10	

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
unsupervised learning, category, schema, triggering, attribute, value, feature, default, variable

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

This research deals with unsupervised learning of categories (UL) and how such learning is affected by the sequencing of training instances. Two general models of UL are described, one based on learning explicit associations between correlated features (associative model), and the other based on creating distinct schemas to represent each category without explicit learning of feature correlations (schema-triggering model). An "attribute listing" paradigm was used as an index of UL in three experiments, each of which manipulated the order in which instances from two different categories were presented and evaluated the effects of this manipulation in terms of the two competing models of UL. Strong evidence was found for the use of a discrete schema-triggering process to learn the categories in these experiments. Moreover, these experiments demonstrate that the attribute listing paradigm can be used to trace learning functions for UL over a series of instances, enabling the future investigation of many independent variables using this task.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL
John F. Tangney, Ph.D.

22b. TELEPHONE *(Include Area Code)*
(202) 767-5021

22c. OFFICE SYMBOL
AFOSR/NL

Previous editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

1 4 FEB 1992

Abstract

This research deals with unsupervised learning of categories and how such learning is affected by the sequencing of training instances. Two general models of unsupervised learning are described, one based on learning explicit associations between correlated features (associative model), and the other based on creating distinct schemas to represent each category without explicit learning of feature correlations (schema-triggering model). An "attribute listing" paradigm was used as an index of unsupervised learning in three experiments, each of which manipulated the order in which instances from two different categories were presented and evaluated the effects of this manipulation in terms of the two competing models of unsupervised learning. Strong evidence was found for the use of a discrete schema-triggering process to learn the categories in these experiments. Moreover, these experiments demonstrate that the attribute listing paradigm can be used to trace learning functions for unsupervised learning over a series of instances, enabling the future investigation of many independent variables using this task.

I. Research Objectives and Summary of Progress

This project's primary goal is to investigate the learning of categories in unsupervised tasks, i.e., when no external tutor is present to provide pre-defined categories and success-related feedback for the learner. A second emphasis is on how category knowledge, once acquired, alters the subsequent interpretation, encoding, and retrieval of individual instances of categories.

During this first year of funding, we have focused our research on two types of task situations. The first task, called attribute listing, involved presenting adult human subjects with a series of instances (in our experiments to date, pictures of fictitious insects) and asking them to write down the distinguishing properties of each instance. The insects varied in type of body, legs, eyes, antennae, wings, etc. As subjects learned which features were common to all category members (defaults), they gradually stopped listing these features and shifted to listing mainly features that varied among members within the category. This listing-pattern occurred because the defaults provided no information relevant to distinguishing among different instances, whereas variable features did provide such discriminative information. The tendency to list variable attributes while omitting defaults provides a measure of category learning which can be traced over trials, i.e., that yield learning functions for the experimental categories.

The second type of task we've investigated consists of presenting subjects with a series of verbal stimuli (e.g., lists of features supposedly possessed by different species of trees) and asking them to study and try to memorize all the features in each list. For example, a particular species of tree might be described as possessing mossy green bark, tall columnar form, deep spreading roots, smooth leaf margins, and so on. Following the study of each list, a series of multiple-choice recognition tests was presented to evaluate subjects' memory for the preceding list. Subjects were allowed to examine only one feature at a time during the study period. A computer recorded how long they spent looking at each feature. As subjects learn which features are defaults (have predictable values) for each category, they spend less time studying these defaults and spend more time studying variable features. This pattern of study times arises because the defaults are predictable to subjects who have learned the experimental categories; these defaults ca be "remembered" (or guessed) easily on the recognition tests without explicit memorization in each individual instance. The decline in study times for default features and corresponding increase for variable features provides an index of unsupervised learning over instances that closely corresponds to that provided by the attribute listing procedure described above.

We have conducted a number of attribute listing experiments over the last year, three of which are described in the detailed report attached. These studies were designed to distinguish between two general theories of unsupervised learning, which we refer to as the *associative* vs. *schema-triggering* models. The associative model assumes that people learn categories by explicitly recording associations between co-occurring features, gradually building up a table of correlations that captures the categorical structure of a given stimulus domain. The schema-triggering model assumes that learners create separate schemas to represent different categories of instances, and thereby capture co-occurrences among features without needing to learn explicit correlational rules or associations. These two models differ in several ways, particularly in their predictions of how the particular sequence in which instances are presented affects the difficulty of learning to distinguish the categories. The detailed report describes several sequencing manipulations designed to discriminate between the two theories of unsupervised learning. Overall, these experiments provided strong support for the existence of a schema-triggering process in unsupervised learning.

With regard to the study time task, our main objective in this first year was to develop and refine the task itself, asking whether it could provide a convergent measure with the attribute listing task for testing theories of unsupervised learning. To this end, we have tried out (in pilot studies) several variations of this task in an attempt to discover which particular experimental arrangements produced the best measurement of unsupervised learning. One of these experiments was analogous to the first attribute listing experiment described in the report below. The results of this study-time experiment replicated those of the earlier attribute listing experiment in essential details. This outcome provided a demonstration of the basic utility of the study time task; it also provided additional evidence for the existence and generality of the schema-triggering process.

In addition to the study time data, the recognition-memory data from this task provide another converging measure of unsupervised learning. Our results show that memory for both the default and variable features of instances improved with category learning, as predicted by the learning model described in our original grant proposal. In fact, memory performance can be traced over trials in this task to reveal category learning curves similar to those from the study times and the attribute listing task. What is particularly interesting about this task is that it makes it possible to observe the simultaneous effects of category learning on instance encoding and retrieval processes. We plan to use this experimental set-up to run several of the planned memory experiments described in the grant proposal during our second year of funding.

II. Planned 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.

2. Clapper, J.P. & Bower, G.H. "Schema-triggering in unsupervised learning." -- This paper reports the results of several attribute listing experiments. To be submitted.

3. Clapper, J.P. & Bower, G.H. "The impact of unsupervised category learning on encoding and remembering instances." -- This paper will describe several study time experiments. To be submitted.

III. Participating Personnel

1. Gordon H. Bower, PI
2. John P. Clapper, Research Associate
3. Terry Nellis, undergraduate Research Assistant

IV. Detailed Report of Attribute Listing Studies

A detailed description of three attribute listing experiments follows.

Schema-Triggering in Unsupervised Learning

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

Several conventional assumptions about stimulus and category representation are presupposed throughout this article. Stimuli will be described in terms of *features*, which are specific values of *attributes*, e.g., size, color, or shape. For example, blue and brown would be possible values of the attribute of eye color in humans. Here, we are concerned with how people learn to distinguish categories based on *correlated* (consistently co-occurring) attribute values. To illustrate, a collection of fruit flies bred in a geneticist's laboratory could be described in terms of several attributes such as size, eye color, wing shape, and so on. If it was then observed that individuals with long wings were also large in size, with red eyes and hairy legs, whereas those with short wings were small with white eyes and hairless legs, these patterns of feature co-occurrences would form an inductive basis for recognizing two distinct categories of fruit flies within that population. Such a characterization of categories in terms of correlated features is consistent with the treatment of Rosch (1975, 1977) and does not imply that the interfeature correlations must be perfect (i.e., that categories be defined by necessary and sufficient features). Since a category would have positive utility so long as some of its features could be predicted with greater-than-chance reliability, the present characterization admits "fuzzy" categories with probabilistic features.

Theories of Unsupervised Learning

Within the framework of these assumptions, the main theoretical objective is to describe how people learn such correlational patterns in real stimulus domains. One theory, which we refer to as the *one-process* or *associative* theory of unsupervised learning, simply assumes that people record associations between all (or some) of the presented features on each learning trial. In this theory, memory can be imagined as a matrix of inter-feature correlations, each of which may be strengthened by experience and weakened by decay and/or interference processes. If some features are consistently correlated in their appearance over many learning trials (instances), the associations among these features will increase in strength relative to those among uncorrelated values. After sufficient training, these correlations would be strongly encoded in memory, and the person could be said to have acquired the category they imply. For example, at this point the person could predict the values of certain attributes given the values of some other attributes, i.e. they could fill in the category's default values.

There are two broad types of feature-association theories. The first class of theories assumes that associations between all presented features are strengthened simultaneously on each trial (e.g., J.A. Anderson, 1977; Rumelhart & Zipser, 1986). We can refer to these models as "matrix autoassociators", since memory is viewed as a matrix of inter-feature associations that are continually updated by new experiences. The second class of theories here are the rule-sampling or hypothesis-testing theories, in which correlational hypotheses are tested sequentially (usually one per trial) against the observed features in each instance (e.g., Billman & Heit, 1988). These rules are strengthened by confirmation and may also be weakened by disconfirmation on a given trial. The main difference between these theories and the matrix models is in whether pairwise associations are strengthened simultaneously or sequentially. For the most part, these differences are not relevant to the research described below, and so they will not be discussed further.

A second framework, which we will refer to as the *two-process* or *schema-triggering* hypothesis, does not require that inter-feature associations be explicitly recorded in memory. Rather, memory needs only record an index of the strength or frequency of each individual attribute value. Adjusting strengths of individual features rather than of pairs of features greatly reduces the amount of information that the learner must keep track of in memory. Dissimilar instances are assimilated to different sets of norms (schemas) in long-term memory (see, e.g., Schank & Abelson, 1977; Rumelhart & Ortony, 1977; Graesser, Woll, Kowalski, & Smith, 1980; Schank, 1982). Thus, inter-feature associations are captured indirectly, by assimilating instances with different sets of correlated values to different schemas in memory, rather than by strengthening associations between feature pairs. By contrast, in the associative models inter-feature associations are explicitly represented in memory whereas categories are present only implicitly.

The information-processing steps in one model of this type (see also Clapper & Bower, 1991) are described below. The learners in this model are assumed to be engaged in unguided exploration of a domain of objects, i.e., learning is unsupervised and learners are simply attending to the features of individual objects without explicitly searching for categories among them. The model provides an example of how schema-triggering plus the strengthing of presented features can describe category learning in unsupervised learning tasks.

1. Categorize the presented stimulus This model assumes that a presented stimulus is automatically classified into the best-fitting category currently available from long-term memory. The category (schema) provides a set of attributes for generating an internal description of the stimulus, plus normative expectations about likely values for each attribute.

2. Evaluate the stimulus features. The features of any stimulus will vary in how well they match the norms of the reference category. The degree of match between an observed attribute value and the category norms for that attribute determine the expectedness or normality of that value (Kahneman & Miller, 1986; Clapper & Bower, 1991). In terms of describing the current instance with respect to its reference category, i.e., distinguishing it from other instances of the same category, the informativeness of a feature is inversely proportional to its expectedness. Since highly expected features are present in many instances of a category, they provide little basis for discriminating among particular instances. By contrast, highly unusual or surprising features are present in relatively few instances of the category, and thus have high utility for distinguishing an instance possessing them from other category members. As successive instances of a novel category are presented, people should learn to discriminate among features on the basis of their discriminative informativeness, ignoring consistently repeated (*default*) values and focusing on surprising or unpredictable information about the stimulus.

3. Encode the instance. The relative informativeness of the different features of an instance determine their attentional allocation or priority of processing. Those features that are most surprising or unusual will receive the lion's share of the subject's attention, while defaults are routinely ignored. The episodic memory representation that results can be thought of as a set or vector of attribute values, each with a specific strength of association to the instance. A feature's strength in this representation would be a direct function of how much attention it received during encoding, which depends in turn on its informativeness with respect to category norms.

4. Modify category norms Learners are assumed to update their category norms after each presented instance. Where these modification are made depends on the degree of fit between the instance and the reference category used to encode it.

4a. Assimilation to an existing category. If the instance does not violate category defaults and cause the subject to invent a new category to accommodate it, then it is assimilated into the previously-activated reference category. The schema for this category is modified by increasing the strength of each presented value in proportion to how much attention it received during encoding, i.e., as a function of its informativeness. Since familiar defaults receive little attention at encoding, their strength in the underlying norms changes little from trial to trial. The strength of more unusual or informative values, by contrast, may be increased greatly due to a single presentation.

4b. Invent a new category. If an instance mismatches the best-fitting reference category in excess of some internal criterion, a new category is "triggered" (i.e., a separate schema is created to represent that category) and the instance is assimilated to this new category. Subsequent instances of this type will then be assimilated to the new category without affecting the norms of the previous category. While the triggering criterion cannot be precisely specified at present, we adopt a hueristic assumption that an instance which violates multiple default values of its reference category will be likely to result in the invention of a new category to handle these discrepancies. Thus, the degree of mismatch or the "surprise value" of a stimulus with respect to prior norms is used as a heuristic strategy for deciding when to invent new categories. This strategy for creating new categories is similar to the "failure-based generalization" of Schank (1982), and the "surprise heuristic" of Holyoak, Nisbet & Thagard (1986).

The schema for the new category is generated by modifying that of the source category to which the instance was first assigned. The model assumes that learners transfer all norms of the source category not specifically violated by the triggering instance to the new category created around that instance. New attribute norms are created only for those attributes whose unusual values triggered the new category. This transfer heuristic ensures that learners need to make the fewest possible changes to their existing norms to handle deviant observations.

Both the schema theory and the explicit learning of inter-feature correlations provide learning methods by which learners might capture the correlational structure of their environment. However, the models differ greatly in their sensitivity to the particular *sequence* in which training instances are presented. In particular, schema-triggering should be vulnerable to aggregation errors early in training, where "aggregation" refers to grouping stimuli that exemplify different correlational patterns into a single category. Such errors could arise because (a) new categories are triggered by violations of strong default expectations, and (b) experience is required for such strong expectations to be formed, so that there may be no strong defaults associated with a category early in training (except those it has inherited from its source category, see above). To illustrate, imagine two categories, A and B, that have contrasting default values along several attributes. If instances of the two categories were presented in a mixed or randomly interspersed sequence from the beginning of training, the triggering hypothesis implies that subjects might often aggregate the two types of instances into a single overgeneralized category. Such aggregation would be an error in the sense that a single category would lose information about feature correlations that would be captured by two separate categories.

By contrast, if several instances of one category were presented before to the first instance of the other, subjects would have time to learn strong defaults for the first category prior to encountering the second. When an instance of the second category was presented, it would then violate default expectations of the first and cause a new schema to be created. Thus, category discrimination should be improved by separating the categories in the training sequence. We show that the strong predicted effect of training sequence on initial discrimination between categories is incompatible with a simple associative model, which expects discrimination to be much less affected by sequencing.

An Index of Unsupervised Learning

We now describe a procedure that can provide information about the course of category acquisition in unsupervised learning tasks. This goal of this method is to trace learning over trials for the default values of each category in a given stimulus set. Specific independent variables can then be evaluated in terms of their effects on these learning functions.

The stimuli in these tasks consist of several attributes, each of which can take on two or more alternative values. Categories in the stimulus sets are defined in terms of correlated attributes values. For example, representing attributes as serial locations in a numerical string, categories could be denoted as Category A = 11111xxx and Category B = 22222xxx, where the numbers (1 or 2) represent default values of particular attributes and the x's indicate that a particular attribute varies independently of the others. The basic task consists of showing subjects a series of such stimuli and asking them to list those features of each stimulus that distinguish it from the other stimuli in the set, while omitting non-distinguishing features from their lists.

Note that if all attributes of the stimuli are uncorrelated, subjects should list the current value of each attribute to distinguish an instance from the other stimuli in the set. By contrast, if the stimulus set is partitioned into categories as above, then for each instance subjects need only list one of its correlated attributes values (or otherwise denote its category membership, to distinguish it from stimuli in the other category), plus the values of the variable (uncorrelated) attributes. There would be no need to list more than one correlated value, since doing so would provide no extra discriminative information either within or between categories. This bias in favor of listing uncorrelated (variable) features while omitting correlated features (category defaults) should evolve gradually over several training trials as successive instances are encountered and subjects learn their consistent properties. Thus, this bias can be used as an index of category learning, i.e., learning should be a monotonic function of the percentage of variables listed minus the percentage of defaults listed.

Importantly, the choice of this learning index is neutral with respect to the associative vs. schema-triggering models. Within the associative model, the difference between default and variable listing results from forming stronger associations in memory between correlated default values than between uncorrelated variables. Within the schema-triggering framework, the difference between defaults and variables lies in their relative strength within the category norms, with default values regarded as less informative than variables due to their greater expectedness.

Experiment 1

The aim of this experiment was to evaluate the attribute listing task as an index of unsupervised learning, as well as its sensitivity to sequence effects. Listing performance over trials was compared in three conditions. In the Blocked condition, the stimuli were partitioned into two categories based on patterns of correlated attribute values. The training sequence was blocked by categories, i.e., a series of instances from one category was presented, followed by a series of instances from the other category. In the Mixed condition, the same stimuli were used as in the Blocked condition, but instances of both categories were randomly interspersed in the training sequence rather than being grouped into separate blocks. In the Control condition, all the attributes of the stimuli varied independently, so that none of the attributes were correlated and the stimulus set was not partitioned into distinct categories.

The first two conditions provided a test of the two models of unsupervised learning described above. A schema-triggering process implies that early aggregation is likely to occur when contrasting categories are presented in a mixed sequence, and so much poorer learning was predicted to occur in the Mixed condition than in the Blocked condition. An associative model could accommodate interference between categories in the Mixed condition by assuming that associative interference results from learning correlations among different values of the same set of attributes. According to this hypothesis, the category presented first in the Blocked sequence should be learned without interference, and thus should be acquired faster than those in the Mixed condition. However, this hypothesis predicts that the second category in the Blocked condition should be learned more slowly than the first, due to proactive interference or negative transfer from the first category on learning and remembering associations between the default values of the second. By contrast, a schema-triggering process predicts that the second category in a Blocked sequence should be learned as rapidly as the first, and no negative transfer from the first category should be observed.

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

Methods

Subjects

The subjects were 30 Stanford University undergraduates participating in partial fulfillment of an Introductory Psychology course requirement.

Procedure

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

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

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

Materials

The stimuli were line drawings of fictitious insects, all of which shared a common "base" structure (e.g., head, thorax, abdomen) plus eight dimensions of variation (attributes), such as wing shape, abdominal markings, eye color, etc. (see Figure 1).

Insert Figure 1 about here

Each attribute had either two or four discrete values (e.g., wings of different shapes, differently colored eyes, and so on), depending on the experimental condition to which it was assigned.

The stimuli shown to a given subject were constructed according to one of two different plans, depending the condition to which that subject was assigned (see Table 1).

Insert Table 1 about here

In two of the three experimental groups, the stimulus set was partitioned into two distinct categories, defined by contrasting sets of correlated attribute values. In these groups five of the eight attributes were binary (two-valued) and their values were perfectly correlated across the instances, such that each instance contained one of two possible sets of correlated values. An instance's category membership was defined by which of these two clusters of correlated values it contained. These values will be referred to as the *default* values of each category.

The remaining three attributes in the Category conditions were four-valued and variable within each category. Two of the four values occurred with equal probability in instances of Category-A, while the other two occurred with equal probability in instances of Category-B. These attributes were uncorrelated within each category, i.e., they varied independently across instances of that category. Within these constraints, eight instances were generated from each category, for a total of sixteen overall.

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

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

Design

There were three between-subjects conditions in this experiment, two of which had correlated values and one of which did not, as explained above. The two correlated conditions employed the same stimuli and differed only in the order in which training instances from the two categories were presented.

In the *Blocked* condition, instances of the A-category were presented in random order for the first sixteen trials, followed by sixteen B-instances (each instance of the two categories was presented twice). After this "training phase", a final test block of eight trials was presented in which four instances from each category were presented together in a mixed sequence. The order of instances in this test block was random, with the restriction that no more than two instances from the same category could occur in a row.

In the *Mixed* condition, the same instances were presented as in the Blocked condition, but in a different order. During the training phase, the 32 A- and B-instances were presented in an intermixed sequence rather than blocked as in the previous condition. Instances from the two categories were presented in random order, with the restriction that no more than three instances from the same category could occur in a row. A final mixed test block of eight instances from the two categories was then presented, the same as that used in the Blocked condition, (i.e., the same specific insect pictures were presented in the same order in both conditions).

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

Counterbalancing

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

Results and Discussion

We begin by discussing the results from the Control condition, since this was intended as a reference group for evaluating category learning observed in the other two conditions. Data from this condition is displayed in Figure 2.

Insert Figure 2 about here

For simplicity, listings over trials for the two psuedo-categories (defined by the base defaults) are separated in this figure, although they occurred together in training.

There are two main results of interest in the Control condition. First, subjects consistently preferred to list four-valued attributes over two-valued attributes; overall, four-valued attributes were listed 19.6% more often than two-valued attributes (t(9) = 3.93, $p < .01$). This indicates that an attribute's discriminative informativeness was perceived as greater when its variability was increased.

Second, there was a significant tendency for subjects to increase their listing of both two- and four-valued attributes over the first few trials, after which listing of both types of attributes remained fairly stable. Within-subjects linear contrasts revealed significantly increasing trends over the first eight instances of both categories for both two- and four-valued attributes (for two-valued attributes, t(9) = 4.47, $p < .01$ for Category A and t(9) = 4.34, $p < .01$ for Category B; for four-valued attributes, t(9) = 2.93, $p < .02$ for Category A and t(9) = 2.09, $p < .10$ for Category B). The tendency for listing to increase during the early trials did not interact the number of values an attribute had, i.e. listing increased about the same amount for both two- and four-valued attributes.

The most likely explanation for this increase assumes that subjects made an incomplete sample of the instances' attributes on the first trial. The initial sample can be thought to correspond to the subjects' hypothesis about which attributes might turn out to be informative (i.e., to differ across instances) over future trials. This hypothesis was then modified over subsequent instances, with attributes added or deleted from subjects' lists based on their observed variability. Subjects in this experiment listed an average of slightly over half of the eight attributes on the first trial. Since these were insufficient to distinguish between later instances, further attributes were added to the list as they were observed to vary across instances.

In the correlated conditions, the four-valued variable attributes displayed a similar pattern of increase over trials as did the corresponding four-valued attributes in the Control condition (compare Figures 2c, 3c, and 4c). The trend was significant over the first eight instances for both categories in the Blocked condition (t(9) = 4.73, $p < .01$ for Category-A and t(9) = 2.97, $p < .02$ for Category-B).

Insert Figure 3 about here

The same was true in the Mixed condition; contrast analyses showed significant linear trends over the first eight instances of both categories (t(9) = 4.79, $p < .001$ for Category-A, t(9) = 3.01, $p < .02$ for Category-B).

Insert Figure 4 about here

In addition, there were no significant differences in overall listings of variable attributes among the three conditions (t(18) = 0.76, $p > .25$ for Blocked vs. Control conditions, t(18) = 1.48, $p > .10$ for Blocked vs. Mixed, and t(18) = .70 for Mixed vs. the Controls).

The defaults showed a very different pattern of results from the variables for the correlated groups. In the Blocked condition, defaults were listed much less often than were the corresponding two-valued attributes in the Control condition (compare Figures 2b and 3b). Averaged over trials, defaults in the Blocked condition were listed with a probability of 17 percent, compared to 66 percent for two-valued attributes in the Control condition (t(18) = 4.70, $p < .001$). Default listing in the Blocked condition also showed a strong learning effect over trials. The proportion of defaults listed declined strongly over the first five A-instances, from 60 percent on the first instance to 7 percent on the fifth; default listing for the B-category decreased from 80 percent on the first B-trial to 17 percent on the sixth. These trends were highly significant by a linear contrast analysis computed over the first six instances of each category (t(9) = 12.78, $p < .001$ for the A-category, and t(9) = 4.14, $p < .01$ for the B-category).

Variable attributes were reported 73% more often than defaults in the Blocked condition (t(9) = 10.54, $p < .001$). The percent listed of variables minus that of defaults on a given trial can be used as a summary index of learning on that trial. These difference scores averaged much higher in the Blocked condition than in the the Control condition (t(18) = 6.44 $p < .001$). The difference scores also showed a clear increasing trend over the first eight instances of each category block; the linear contrasts were significant for both the A-category (t(9) = 9.01, $p < .001$) and the B-category (t(9) = 4.58, $p < .01$). These results demonstrate that subjects learned to discriminate among attributes based on their predictability during the training blocks. The larger difference between two- and four-valued attributes in this condition compared to the Control group was due to the presence of correlational patterns in the stimuli of the present condition. The pattern of decreasing default (and increasing variable) listings make it possible to trace this learning as it increases over instances.

The pattern of responses during the test block was similar to that of the immediately preceding trials. Compared to corresponding trials in the Control condition, listing of defaults during the test block was significantly lower in the Blocked condition (t(18) = 3.66, $p < .01$), that of variables about the same (t(18) = 0.95, $p > .25$), and the differences between them (learning scores) greater (t(18) = 4.23, $p < .01$). The fact that higher listing of variables than defaults continued during this block, which presented instances of both categories in random order, indicates that the earlier biases in subjects' listings were not a mere artifact of presenting instances of the same category together in the training sequence. That is, the increase in subjects' learning scores reflects the acquisition of stable categories rather than local habituation to a series of repeated values.

Subjects' attribute listings also provide strong evidence for learning in the Mixed condition (see Figure 4). Default values were listed 36 percent less often in this condition than the corresponding two-valued attributes in the Control condition (t(18) = 3.54, $p < .001$). Since variable listing was about the same in the two conditions, difference scores were also higher in this condition than for the controls (52 vs. 20 percent, t(18) = 3.54, $p < .002$). These results indicate that subjects in the Mixed condition discriminated more between defaults and variables in their listings than could be explained by a mere preference in favor of listing four-valued rather than two-valued attributes. Rather, the additional bias indicates that subjects' listings were affected by the feature correlations in the Mixed condition.

A comparison of Figure 3b and Figure 4b suggests that learning occurred much more rapidly in the Blocked condition than in the Mixed condition. In fact, no default learning appears to have occurred in the Mixed condition until after the first five or six instances of each category. Prior to this, listings remained at a fairly constant level, and neither default listings nor difference scores differed significantly from the same trials in the Control condition. A linear contrast analysis showed no decrease in default listing over the first six trials of either category (t(9) = 0.99, $p > .25$ for Category A and t(9) = 1.05, $p >$ .15 for Category B). Listing of defaults began to decrease in the trials following this, although the linear trend for default listing did not reach conventional levels of statistical reliability over trials seven to sixteen for either category (t(9) = 1.57 for Category A and t(9) = 1.58 for Category B; $p > .10$ for both tests). The difference scores were apparently a more sensitive indicator of learning in this condition, and showed significant increases over the first ten trials for both categories (t(9) = 6.52, $p < .001$ for Category A and t(9) = 2.52, $p <$ 05 for Category B).

Direct statistical comparisons between the Blocked and Mixed conditions support the conclusion that learning occurred more rapidly in the Blocked condition. The mean proportion of default values listed was greater in the Blocked condition by the third instance of Category-A and by the second instance of Category-B. In addition, learning in the Blocked condition appeared to be complete in less than five instances for both categories, whereas default listings in the Mixed condition required much longer to reach their minimum level. Overall, default listings during the training phase were significantly lower in the Blocked condition for Category-A (t(18) = 2.26, $p < .05$), although not for Category-B (t(18) = 0.70, $p > .25$).

The difference scores appear to have been a more sensitive indicator of sequence effects in this experiment, probably because variables were listed slightly more often in the Blocked than the Mixed condition (a non-significant difference, t(18) = 1.48, $p > .10$). Difference scores were higher in the Blocked condition for the first eight instances of both Category-A (t(18) = 3.74, $p < .002$) and Category-B (t(18) = 2.35, $p < .05$), and marginally greater for the second eight instances of Category-A (t(18) = 1.96, $p < .10$). Pooled over the 32 training trials, difference scores were significantly greater in the Blocked than the Unblocked condition (t(18) = 2.47, $p < .05$). For the final test block, there was no significant difference between Blocked and Unblocked conditions for either difference scores (t(18) = 0.72, $p > .25$) or default listings alone (t(18) = 0.03). This suggests that although learning occurred more rapidly in the Blocked condition, subjects in the Mixed condition caught up by the end of the experiment.

The schema-triggering hypothesis provides a plausible explanation for the slower learning that occurred in the Mixed condition. This hypothesis implies that subjects would be likely to aggregate both types of instances into a single category when they are presented together early in training, thus failing to capture the correlational structure of the stimulus set. This should occur because subjects in the Mixed condition would have less time to learn strong defaults for one category before seeing instances of the other. Due to the lack of strong default expectations, the novel stimulus would be less likely to trigger the formation of a separate schema and would be more likely to be aggregated together with previous instances into a single overall category. It might be difficult for subjects to "unlearn" this aggregated category and acquire the correct category-level discriminations. Assuming that some subjects discriminated the categories correctly from the start of training (triggering a new category upon seeing the first discrepant stimulus), some aggregated the categories together at first but later overcame this initial error, and that some never unlearned their initial overaggregation, the averaged data might match the pattern of gradually increasing learning observed in this condition. (A process by which the schema-triggering model could correct for initial errors of overaggregation will be described in more detail in the General Discussion.

An associative model could explain negative transfer in the Mixed condition as due to associative interference in learning correlations among different pairs of values from the same set of attributes. A strong interference process could explain why Category A learning was reduced by interspersing B instances in the training sequence in the Mixed condition. However, such an interference process would imply that prior learning of Category A in the Blocked condition should have interfered with subsequent learning of Category B, as well. The data show no such negative transfer; if anything, the second category was learned slightly faster than the first in this group. The associative model provides no obvious explanation for why A instances would interfere with B learning when interspersed with B instances in the Mixed condition, but not when presented first as in the Blocked condition.

In sum, the results of this experiment suggest that the attribute listing task can be productively used as an index of unsupervised learning for both Blocked and Mixed training sequences. Moreover, the patterns of transfer revealed in comparing these two groups provide evidence that are difficult to accommodate within the associative model but are readily explained by schema-triggering.

Interestingly, the base defaults behaved somewhat differently in this experiment than did the other defaults, and the schema-triggering idea can also accommodate these differences. Recall that the base defaults were judged to be the most salient attributes of the insect stimuli, and it was considered likely that subjects would tend to list these particular attributes when they wished to indicate an instance's category membership. To illustrate, people should prefer to describe the categories as "broad-winged" versus "narrow-winged" than as, say, "black-eyed" versus "white-eyed", because wings were more physically prominent than eyes in these stimuli. Consistent with this, base defaults were listed more often whenever subjects would be expected to want to indicate an instance's category membership. For example, when a long series of instances from the same category is presented in sequence, the category membership of each could be readily inferred on the basis of this local context. But when instances are presented in mixed sequence and category membership cannot be inferred from local context, subjects could indicate it by listing the most physically prominent default (i.e., wings or body snape) as a proxy for the category.

Consistent with this account, higher listings were observed for base defaults in the mixed test block of the Blocked condition than in the last eight trials of preceding same-category training blocks (t(9) = 2.48, p < .05). No such increase occurred for either variables (t(9) = 1.00, p >.25) or regular defaults (t(9) = 1.54, p > .10). In other respects the base defaults behaved like the regular defaults in the Blocked condition, decreasing strongly over the first six instances of each category (t(9) = 2.83, p < .05 for Category A and t(9) = 6.85, p < .001 for Category B). Base defaults stayed fairly constant throughout the task in the Mixed condition, showing no significant decreasing trends. Any subjects that learned the categories in the Mixed condition would need to explicitly indicate the category membership of each instance, since this could not be inferred from context.

Experiment 2

The aim of this experiment was to provide further evidence to discriminate between the associative and schema-triggering theories. One difference between them is that the associative theory expects learning of a category to increase monotonically with the number of instances presented, i.e., that adding to the number of A instances present in a training sequence should always increase, or at least not decrease, final A learning. By contrast, the schema-triggering theory predicts that in certain situations learning could actually be decreased by increasing the number of instances presented from a given category. This could occur if the added instance interfered with initially forming distinct schemas for the two categories, and caused them to be aggregated together into a single category instead. The present experiment aims to provide an empirical demonstration of this prediction of the schema-triggering process.

A second difference is that the associative theory expects transfer or interference effects between contrast categories to be consistent regardless of how the categories are sequenced or the number of instances presented from each. For example, if learning of two categories is reduced when they are presented together in a mixed training sequence, as in Experiment 1, then there should also be negative transfer from Category A on learning Category B in a blocked sequence. Similarly, if presenting four A instances prior to seeing any Bs interferes with B learning, then increasing that number to eight A instances should, if anything, increase the degree of proactive interference on B. By contrast, schema-triggering implies that the direction of transfer (positive or negative) could in some cases be reversed by manipulations of instance sequencing. Thus, a second aim of this experiment was to test some transfer predictions of the schema-triggering hypothesis that cannot readily be accommodated within a simple associative model.

In the following experiment, A and B instances were presented in two different conditions. In one, a "pretraining" block of eight A-instances was followed by a "test" block of twelve A-instances and twelve B-instances presented in mixed sequence. In the other, a mixed pretraining block of four A-instances and four B-instances was followed by the same test block as in the first condition. In the first condition, called the *Contrast* condition, the schema model predicts that subjects would learn strong A-defaults prior to encountering their first B-instance. Thus, they should easily notice the contrast between the two categories when they encounter this instance, invent a new category to accommodate it, and rapidly learn new default values for this category. Moreover, encountering the B-instances should not cause subjects to unlearn or discard the prior A-defaults, i.e., listing of A-defaults should not increase appreciably during the mixed test block. The schema model predicts that the triggering instance should be assimilated to the new category it causes the learner to invent, not to the "source" category to which it was initially assigned. In the second condition, called the *Practice* condition, learning should be reduced because subjects will tend to aggregate the two types of instances into a single category, which ignores the correlational structure of the stimulus set.

Although the schema-triggering theory predicts better learning of B-defaults in the Contrast condition, a larger number of B-instances actually occur in the Practice condition. A total of four B-instances are presented during the pretraining block in the Practice condition, whereas no B-instances occur prior to the test block in the Contrast condition. The associative theory clearly expects better learning of Category B in the Practice condition, since the inter-feature associations among the B-defaults receive more practice (repetitions in different instances) in that condition. Moreover, while the triggering theory predicts that increasing the number of B instances in the pretraining block from zero to four should interfere with later learning of Category A, increasing the number of A instances from four to eight is expected to have the opposite effect on later B-learning. The associative model cannot handle this complex dependence of transfer effects on the sequencing and number of instances presented from each category. If the predicted results were obtained, they would provide strong evidence for the existence of a schema-triggering process in unsupervised learning.

Method

Subjects

The subjects were 40 undergraduate students of San Jose State University participating in partial fulfillment of an Introductory Psychology course requirement.

Procedure

Subjects were tested in groups for a single session of 30 to 45 minutes. The procedure was the same in most respects as in Experiment 1. The training instances were realistic line drawings of fictitious insects, presented in booklets similar to those used in Experiment 1. The same instructions were used as in Experiment 1, i.e., subjects were instructed to list the distinctive properties of each individual insect, where distinctive properties were those that would be useful for distinguishing the current instance from others of the same general type, for example, on a later multiple-choice recognition test. Subjects were instructed to list only those properties that would be useful for identifying an insect on such a test, and to omit non-distinguishing properties even if they were highly prominent or noticeable.

Materials

The same type of pictorial insect stimuli were used as in Experiment 1. These stimuli all shared a common "base" structure (e.g., head, thorax, abdomen) plus eight dimensions of variation (attributes), such as wing shape, abdominal markings, eye color, etc. Each attribute had either two or four discrete values (e.g., wings of different shapes, different colored eyes, and so on), depending on the experimental condition to which it was assigned. Five of the eight attributes had two values, and these values were correlated across instances, such that the set was partitioned into two distinct categories defined by contrasting sets of default attribute values (see Table 2).

Insert Table 2 about here

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

Design

There were two between-subjects conditions in this experiment.

In the *Contrast* condition, instances of the A-category only were presented for the first eight trials, followed by a mixed block of twelve A-instances and twelve B-instances. The first block of eight trials will be referred to as the *pretraining* block, while the second block of 24 instances will be referred to as the *test* block. The first instance of the test block was always a member of Category B. Instances of the two categories were presented in a randomly ordered, intermixed sequence, with the constraint that no more than three instances from the same category were allowed to appear in a row.

In the *Practice* condition, the eight instances from the pretraining block consisted of four As and four Bs, rather than eight As as in the previous condition. The four instances from each category were selected so that both values of each variable attribute occurred twice, and none of the variable attributes was correlated with any of the others. They were presented in a random order, with the restrictions that the first instance was a member of Category-A and that no more than two instances from the same category could occur in sequence. The same 24-instance test block was used as in the Contrast condition.

Note that the only difference between the two conditions is that in the Practice condition four B-instances were substituted for four A-instances presented in the Contrast condition.

Counterbalancing

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

Results and Discussion

The Practice condition in this experiment was essentially a replication of the Mixed condition of Experiment 1. The results are displayed in Figure 5.

Insert Figure 5 about here

---

[1] Compared to the Mixed condition from Experiment 1, somewhat less learning seems to have occurred in the present condition. Default listings appear to decrease slightly over the course of the experiment, but the decreasing trends are not statistically significant by linear contrasts conducted over various intervals of trials. Nor were default listings averaged over the eight trials of the pretraining block lower than those averaged over the twelve subsequent A-trials (t(17) = 0.07, $p > .50$) or B-trials (t(17) = 1.51, $p > .10$). For the base defaults, a significant difference between early and late trials was obtained in Category A (t(17) = 2.32, $p < .05$), but not in Category B (t(17) = 1.16, $p > .10$).

Turning to the difference scores (listing of defaults subtracted from that of variables) a significant increase occurred over the first four instances of Category A (t(17) = 3.71, $p < .01$) and the corresponding instances of Category B (t(17) = 4.26, $p < .001$). Some of this increase was due to increased listing of the variable attributes of both categories during the same trials. This increase was significant by a contrast analysis for linear trends over the first four instances of Category A (t(17) = 3.35, $p < .01$) and the first four instances of Category B (t(17) = 5.31, $p < .001$). Following the first two pretraining trials, all difference scores were positive (i.e , variables were listed more often than defaults throughout most of the experiment).

The apparent learning effects in this condition appeared smaller than those from the corresponding condition of Experiment 1. However, fewer trials were used in the present experiment (32 instead of 40), and a different subject population was sampled (students of San Jose State University instead of Stanford University). In addition, it is useful to compare the present results to those of the Control condition from Experiment 1, in which correlated default values were lacking. In that condition, subjects significantly *increased* their listing of two-valued attributes over the first eight to ten trials, as they became aware that these attribute varied independently over instances and thus were informative for the listing task. Seen in this light, the slight decrease in default listing observed in the present experiment probably indicates some real learning of these defaults.

Without an uncorrelated control condition, it cannot be conclusively demonstrated that default learning occurred in the Practice condition of the present experiment. However, in this experiment we were mainly concerned with *differences* in learning between the Practice and Contrast conditions. As expected, the pattern of results from the Contrast condition differed sharply from those of the Practice condition (see Figure 6).

Insert Figure 6 about here

Here, all the instances presented during the pretraining block were from Category A. The listing of both A-defaults and base defaults decreased rapidly during this block, from a high of about 41 percent (for defaults) on the first trial to about 6 percent on the eighth trial. The linear trend over this interval was significant at the .001 level for both defaults (t(16) = 4.60) and base defaults (t(16) = 5.14). During the same trials subjects increased their listing of variable attributes from 35 to 73 percent (t(16) = 3.91, $p <$ .01). A significant linear trend was also observed for the difference scores over this interval (t(16) = 5.23, $p < .001$). In sum, the same rapid learning of A-norms that occurred in the Blocked condition of Experiment 1 was observed in the Contrast condition of the present experiment.

Following the pretraining block, a large increase in default listing occurred when the first B-instance was presented, from 6 percent on the previous A-trial to 53 percent on the first B-trial (t(16) = 5.37, $p < .001$). The same effect was apparent in the listing of base defaults (t(16) = 8.17, $p < .001$). Following this initial reaction, listing of B-defaults decreased rapidly on subsequent trials. Most of this decrease occurred between the first and second B-instance (t(16) = 5.10, $p < .001$), with much less change in the learning function occurring thereafter. The same was true of the base defaults (t(16) = 3.06, $p <$ .01), although listing of these attributes remained higher than those of the defaults; his difference probably reflects subjects' continuing use of these highly prominent features to indicate the category membership of each instance during the mixed test block. Overall, default learning during this block appeared at least as rapid as the learning of A-defaults that had occurred during the pretraining block, and showed no evidence of interference from the preceding block of A-instances.

Following the first B-instance in the test block, listing of A-defaults also increased by a small amount (about 12 percent); this increase was statistically significant at the .05 level (t(16) = 2.40). This elevated reponding continued on the second B-instance of the test block, and then tapered off over the new few trials (see Figure 6). Despite their temporary elevation, listing of A-defaults on this trial was still substantially less than that of the B-defaults on the first B-trial (by about 32 percent, t(16) = 3.13, $p <$ .01). This pattern of results seems to indicate that presenting the first B-instance did have some effect on the default norms of Category-A, contrary to our original predictions. However, subsequent B-instances apparently did not affect A-norms (i.e., they did not increase listing of A-defaults) suggesting that they were assimilated only to the newly-invented schema for Category B, as predicted by the schema theory.

An important prediction of the schema-triggering theory was that B-defaults should be learned more rapidly following a block of pure A-instances than following a mixed block composed of both A and B-instances. This result was predicted due to the greater initial learning of A-defaults that would occur in the first condition, which favors the triggering of a new schema at the first B-instance. The result was expected despite the larger number of B-instances presented to subjects in the Practice condition, i.e., in spite of the fact that the inter-feature associations of Category B received more repetition in that condition. Consistent with this prediction, B-defaults were learned much more rapidly and completely in the Contrast condition than in the Practice condition. On the first B-trial in the test block of the Contrast condition, default listing was significantly higher than on the corresponding trial of the Practice condition (t(33) = 2.05, $p < .05$). This reflects the greater surprisingness of those attributes in that condition; the B-values would have been considered default violations by subjects in the Contrast condition, while many subjects in the Practice condition would have merely regarded them as routine values of familiar variable attributes. Following the first B-trial, default listing for the next eleven B-instances was lower in the Contrast condition than in the Practice condition (by an average of 18 percent, t(33) = 2.76, $p < .01$).

The pattern for base defaults was similar to that for defaults, except that listing of these attributes did not show as much decrease over trials as did other defaults. Base default listing was significantly higher in the Contrast condition for the first B-instance of the test block (t(33) = 2.16, $p < .05$). Following this, however, there was no significant difference between the two groups in their listing of these attributes (t(33) = 0.54). This probably reflects subjects' tendency to continue listing base defaults to indicate the instances' category membership in the mixed test block of the Contrast condition.

The listing of variable attributes was approximately the same in the two groups (t(33) = 1.04, p > .10). Thus, the pattern of results for the difference scores simply mirrored those for the defaults, and will not be discussed separately.

Although subjects in the Practice condition saw a larger number of B-instances than did subjects in the Contrast condition, learning of B-defaults in that condition suffered from negative transfer due to the four preceding A instances (compared to the learning that would have occurred had only B-instances been presented during the pretraining block). Importantly, this interference cannot have occurred at the level of inter-feature associations (or explicit correlational rules), because in that case the amount of interference from the A-category should have increased directly with the number of A-instances in the pretraining block, and thus have been greater in the Contrast condition than in the Practice condition. By contrast, increasing the number of A instances from four to eight eliminated their interference on subsequent B-learning. The interference in the Practice condition is explained by the schema-triggering theory as due to inadequate learning of A-defaults prior to encountering the first B-instance, causing subjects to aggregate both types of instances into a single category. Contrary to an associative interference hypothesis, increasing the number of A-instances can either facilitate or interfere with later learning of B-defaults, depending on how the manipulation affects the schema-triggering process (i.e., the probability that triggering will occur at any given point in the sequence).

While our results show that the learning of B-norms was apparently unimpaired by prior A-learning in the Contrast condition, there did appear to be a temporary effect on A-norms due to presenting the first B-instance, i.e., listing of A-defaults increased for several instances following the presentation of the first B-instance. By contrast, we expected that the B-instance would trigger the invention of a new category (which apparently occurred), and that the instance would be assimilated *only* to the new category and would not affect listings for later instances of the source category (A). One explanation for the increase is that while the first B-instance triggered a new category as expected, the instance could have been assimilated *both* to this new category and to Category-A. The new category would then provide a better match to subsequent B-instances than would Category A, so for these later instances only the new B category would be evoked. Meanwhile, the A norms would gradually return to previous levels as subsequent A-instances were assimilated. The only difference between this account and the schema-triggering model presented above is that it assumes that instances are always assimilated to the category to which they were first assigned. If an instance is also sufficiently novel to trigger a new category, then it will be assimilated to that new category as well.

Experiment 3

This experiment was a modification of Experiment 2 designed to further investigate schema-triggering in unsupervised learning. The patterns of transfer (i.e., how the two categories interfere with or facilitate each other's learning) in this experiment were expected to provide further evidence requiring the existence of a schema-triggering process. In particular, the present experiment investigated the effect of initially over-aggregating two contrast categories into a single class on subjects' ability to eventually acquire the correct category-level discriminations.

All the conditions in this experiment resembled the Contrast condition of Experiment 2, except that the series of same-category instances in the pretraining block was preceded by a single instance from the contrasting category. In the Contrast condition of Experiment 2, eight instances of Category A were presented in a row prior to a mixed block consisting of both A- and B-instances. These eight instances were sufficient for most subjects to learn strong A-defaults prior to encountering the first B-instance, causing a new schema to be triggered upon seeing the B-instance. In the present experiment, rather than presenting all A-instances during the pretraining block, a single A-instance was presented during the first trial followed by a series of B-instances (by convention, we always refer to the first-presented category in the training sequence as Category A). The main independent variable in this experiment was the number of B-instances that followed the first A-instance in the pretraining block; one group of subjects had four B-instances in this series, a second group had eight, and a third group had twelve B-instances. Following this pretraining block, a mixed block of both A- and B-instances, similar to that of Experiment 2, was presented for the next thirteen trials.

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

By contrast, if fewer B-instances were presented prior to the second A, the probability of triggering a new category should be reduced. This reduction would result from the relatively high residual strengths of the A-values in the aggregate norms, which would lessen the perceived disparity between those norms and the features of the second A-instance. If subjects failed to dis-aggregate the two categories (i.e., did not create a separate schema for Category A), then their attribute listings in the mixed block should show reduced learning of the default values of both categories. In sum, the existence of a schema-triggering process would imply that increasing the number of B-instances in the pretraining series would increase increase the subsequent learning of both categories.

A simple associative model lacking a schema-triggering process would predict a somewhat different pattern of results. Such models expect that increasing the number of B-instances in pretraining should increase later B-learning, consistent with the triggering hypothesis. However, the associative theory expects that this manipulation would also *decrease* later A-learning due to negative transfer at the level of inter-feature associations. In general, the associative theory predicts that transfer effects will be consistent in strength and direction (positive or negative). For example, if presenting a single A-instance interferes with learning the defaults of subsequent B-instances, as both theories predict, then presenting four to twelve B-instances should greatly reduce default learning in subsequent As.

In addition to providing another test of schema-triggering, the present experiment may also provide an estimate of how many instances from one a category must be presented to overcome initial aggregation with its contrast category, at least within the present attribute listing set-up. In the Blocked condition of Experiment 1 and the Contrast condition of Experiment 2, both of which were favorable for schema-triggering and rapid default learning, about three to five instances were required to fully learn a category's defaults. The present situation should be less favorable for rapid category learning, due to aggregation of the two categories at the start of training. Thus, a larger number of instances should be required to learn the category to asymptote and cause triggering when a contrasting stimulus is presented. Presumably, the degree of learning at the end of the pretraining block (the difference between variable and default listings) should predict learning of both categories during the following mixed block.

Method

Subjects

The subjects were 36 undergraduate students of Stanford University participating in partial fulfillment of an Introductory Psychology course requirement.

Procedure

The procedures for this experiment were essentially the same of those of the previous two experiments. Subjects were tested for a single half-hour session in groups of eight to ten. They were given test booklets similar to those of the other experiments, and allowed to complete the listing task at their own pace. The listing instructions were identical to those used in Experiments 1 and 2.

Materials and Design

The stimuli in this experiment were the same pictorial insect stimuli used in the last two experiments. These were divided into categories on the same basis as the stimuli in Experiment 2. The stimulus set was partitioned into categories on the basis of perfectly correlated values on five binary attributes. The remaining three attributes varied independently over two values, different for the two categories. The design shown in Table 2 for Experiment 2 was also true for the present study.

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

Each of the sixteen possible instances from this set was presented at least once in this experiment, and instances were selected for a second or third presentation such that each value of the variable attributes appeared an equal number of times. As in Experiment 2, two different stimulus sets were generated such that assignment of default or variable status to a given attribute was balanced across the experiment as a whole; this balancing was depicted in Table 2. For both of these stimulus sets, booklets were constructed such that one category of insects took on the role of Category A (i.e., was presented first) for a given group of subjects while another group received booklets in which the other category was presented first. Crossing these two balancing factors (the stimulus set used and the order in which categories were presented) with the three levels of the N variable (number of B-instances in the pretraining series) yielded a total of twelve groups. Three subjects were randomly assigned to each group, for a total of 36 subjects in this experiment.

Results and Discussion

The main data for this experiment (listing of variables minus that of defaults for the three conditions) are shown in Figure 7.

Insert Figure 7 about here

As Figure 7 shows, there is evidence for learning of both categories in all three conditions of this experiment. Starting with Category B, default listing decreased over trials during the pretraining block in all conditions. The decreasing linear trends in default listing were significant over the N trials in the block for the N=4 and N=12 conditions (t(11) = 3.15, *p* < .01 and t(11) = 3.20, *p* < .01, respectively), but not over the block as a whole for the N=8 condition (t(11) = 1.22, *p* > .10). However, the decrease was significant over the first 3 trials of pretraining for the N=8 condition (t(11) = 3.00, *p* < .02).

The difference scores increased significantly during the pretraining block for Category B in all three conditions (t(11) = 4.66, *p* < .001 for N=4; (t(11) = 6.70, *p* < .001 for N=8; and (t(11) = 5.64, *p* < .001 for N=12). The difference scores show more learning than the defaults because they count both the increased listing of variables and the decreased listing of defaults that occurred during this block. This increase was significant in all three conditions (t(11) = 3.99, *p* < .01 for N=4; t(11) = 6.74, *p* < .001 for N=8; t(11) = 4.91, *p* < .001 for N=12).

Default listings tended to increase somewhat (and learning scores to decrease correspondingly) for B-instances during the following test block . Comparing average default listing for the test block with that of the last two B-instances in the pretraining block, a marginal increase in listings was observed in N=4 (t(11) = 1.90, *p* < .10), a significant increase in N=12 (t(11) = 2.68, *p* < .05), and a non-significant increase in N=8 (t(11) = 1.42, *p* > .10). However, the listing of defaults remained far below that of variables during this block (t(11) = 5.64, *p* < .001 for N=4; t(11) = 5.09, *p* < .001 for N=8; and t(11) = 7.76, *p* < .001 for N=12). This indicates that the learning of B-defaults that occurred during pretraining transferred to the test block, and was not merely due to temporary habituation to a series of repeated values.

The listing of A defaults also declined significantly over trials in condition N=12 (t(11) = 2.86, *p* < .02), nearly significantly in N=4 (t(11) = 1.76, *p* < .12), and non-significantly in N=8 (t(11) = 0.19, *p* > .20). Variable listing increased significantly for Category A in all three conditions, but only over the first three instances (t(11) = 3.38, *p* < .01 for N=4; t(11) = 5.70, *p* < .001 for N=8; and t(11) = 5.63, *p* < .001 for N=12). Difference scores also increased over the eight A-instances in all three conditions (t(11) = 4.68, *p* < .001 for N=4; t(11) = 6.05, *p* < .001 for N=8; and t(11) = 7.54, *p* < .001 for N=12).

The primary aim of this experiment was to compare learning among the three groups, and show that a longer series of B-instances in the pretraining block would cause better learning of both categories in the test block. As in Experiment 2, the lack of an uncorrelated control condition in this experiment weakened the within-groups results as evidence for category learning in the individual conditions. That is, it is possible that the higher listing of variables than defaults in some conditions simply reflected subjects' preference to list four-valued rather than two-valued attributes. However, comparing data from the present study to the uncorrelated Control condition in Experiment 1 supports the conclusion that correlational learning probably occurred in all three groups in this experiment. Recall that listing of uncorrelated attributes with both two and four values increased significantly over trials in Experiment 1, in contrast to the gradually decreasing pattern of default listing in the present experiment. If subjects in this experiment were listing variables more often than defaults only because the former had more possible values, listing of both types of attributes should have increased over trials as in Experiment 1. The fact that default listing tended to decrease in this experiment argues for real category learning in the present experiment.

We now turn to the between-groups analyses and to the specific tests of our theoretical hypotheses. The main prediction derived from the schema-triggering hypothesis was that increasing the number of B instances during the pretraining block should increase learning of both categories in the following mixed block. This was expected because increasing the number of B instances should increase the relative strength of B-values in the aggregated norms, while decreasing the relative strength of residual A-values from the first trial. This, in turn, should increase the probability of triggering a new schema when the next A-instance was encountered, because the A-values would have low strengths in the aggregated norms and hence should appear relatively surprising with respect to those norms. Once the categories were dis-aggregated by this triggering, default learning could occur rapidly for each.

These expectations were largely borne out, with one qualification. We originally expected that the degree of learning would vary monotonically with the length of the pretraining block, i.e., that learning would be greater in the N=12 group than for N=8, and greater in N=8 than in N=4. While learning did tend to be higher in condition N=12 than in the other two conditions, learning in N=8 was not greater than that in N=4; if anything, it tended to be slightly less. Thus, increasing the pretraining block from N=4 to N=8 did not improve learning, but increasing it to N=12 did.

Turning to statistical comparisons, default listing for category A was significantly less in N=12 than in N=8, t(22) = 2.23, p < .05. The difference was also significant for the difference scores, t(22) = 2.85, p < .01. Although A-defaults were listed less often in N=12 than in N=4 (by 7.6%), and the learning scores are higher in N=12 (by 10%), neither of these comparisons attained conventional levels of statistical reliability (t(22) = 0.76 and t(22) = 0.95, respectively). Learning appeared to be somewhat higher in N=4 than in N=8, as noted previously, but these differences also failed to reach statistical significance (t(22) = 1.40 for defaults and t(22) = 1.56 for variables, both p-values > .10). When the data from conditions N=4 and N=8 were pooled, the comparison between difference scores in this combined condition and in N=12 was marginally significant (t(34) = 2.02, *p* < .10).

Comparisons of Category B learning showed a similar ordering of conditions as did those of Category A. Within the pretraining block, learning appeared greater in N=12 than in N=8 and N=4, but not greater in N=8 than in N=4. Comparing the final trial of the pretraining block in each condition, default listing was significantly less in N=12 than in N=8 (t(22) = 2.27, *p* < .05) and in N=4 (t(22) = 2.24, *p* < .05), but there was no difference between N=8 and N=4 (t(22) = 0.81, *p* > .25). The difference scores showed the same ordering of learning, although the effects were somewhat weaker than those shown by the defaults. Difference scores on the last pretraining trial were marginally greater in N=12 than in N=8 (t(22) = 2.06, *p* < .10), non-significantly greater in N=12 than in N=4 (t(22) = 1.47, *p* > .10), with no significant difference between N=8 and N=4 (t(22) = 0.75, *p* > .20).

Turning to the test block, B learning was again higher in N=12 and lower in the other two conditions. Averaged over the six B-instances in the test block, these effects were statistically significant for the difference scores but not for default listings alone. Difference scores for N=12 exceeded those of N=8 by 27%, a significant difference (t(22) = 2.14, p < .05). In addition, difference scores were 22% higher in N=12 than in N=4, a marginally significant effect (t(22) = 1.73, p < .10). No significant difference was obtained between N=4 and N=8 (t(22) = 0.41). When N=4 and N=8 were pooled together into a single condition, difference scores in this condition were significantly less than those in the N=12 condition (t(34) = 2.28, *p* < .05).

Although learning in the N=12 condition was higher than that of the other two groups in this experiment, it still did not appear as high as in the test block of the Contrast condition from Experiment 2. In that condition, all instances presented during the pretraining block were members of the same category, and thus there was no initial aggregation of categories for subjects to unlearn. Although listing of defaults rose during the first part of the test block for both categories in that experiment, they declined to near their original levels very quickly, within two or three trials. Although the validity of direct comparisons between these two experiments is questionable because different subject populations were sampled, it appears that less learning occurred during the test block of Experiment 3, even in the N=12 group. Apparently, a single A-instance presented before as many as twelve B-instances was enough to interfere somewhat with later discrimination of the two categories. This is a surprisingly strong negative effect, and additional research is needed to further explicate these powerful transfer effects and the conditions under which they are likely to occur.

Overall, these results are consistent with the schema-triggering hypothesis but not with the one-process associative model, since the latter cannot account for the increased A-learning that occurred due to increasing the number of preceding B-instances. However, neither theory provides any simple explanation for why learning should have been less in N=8 than in N=4. Perhaps the most plausible interpretation of these results is that no real differences existed between N=4 and N=8, only between these two conditions and N=12. Although N=8 appeared to show slightly less learning in some comparisons than N=4, none of these comparisons were statistically significant. Moreover, if learning is traced over the first eight trials of the N=12 group, and compared to learning observed during the corresponding trials of the N=8 group, more learning seemed to occur in N=12. For example, the second four instances in this block showed significantly lower default listings than the first four in N=12 (t(11) = 2.51, *p* < .05), but not in N=8 (t(11) = 0.90, *p* > .20). Thus, it may be that the poor learning observed in condition N=8 was to some degree a random effect.

Nevertheless, the present results raise the possibility of a "threshold" effect for triggering new categories, which in this particular experiment occurred between the eighth and the twelfth B-instances. What would be the implications of such a threshold for the adequacy schema-triggering as a general explanation of learning from both blocked and mixed sequences? Possibly, subjects might use schema-triggering to speed their category learning when conditions are favorable for such triggering to occur, but also track feature correlations to learn categories more slowly when conditions for triggering are unfavorable. This issue will be discussed in more detail below.

General Discussion

These three experiments showed powerful effects of the sequencing of training instances on unsupervised learning. In some cases, learning of a category was improved the greater the number of instances presented from that category. For example, in Experiment 1 learning of both categories increased with number of instances for both Blocked and Mixed training sequences. In Experiment 2, adding A-instances to the pretraining block (changing it from four As and four Bs to eight As) increased the level of A-learning in the subsequent test block. And in Experiment 3, adding B instances to the pretraining block (increasing from four to twelve) improved B-default learning in the test block. Sometimes, however, learning was actually reduced by increasing the number of instances presented from a given category. This seemingly paradoxical effect occurred in Experiment 2 when adding B-instances to the pretraining block (changing it from a pure block of A-instances to a mixed block containing instances of both categories) decreased later B-learning in the test block.

A similar interaction with sequencing was observed for transfer effects. For example, Category A learning in Experiment 1 was greatly reduced by interspersing the A training instances with members of Category B (Blocked vs Mixed conditions), an example of negative transfer. In Experiment 2, both negative and positive transfer between the categories was observed. Adding A-instances to change a mixed pretraining block to a pure A block improved subsequent learning of category B, a demonstration of positive transfer. By contrast, adding B-instances to change a pure A block to a mixed block of As and Bs had a strong negative effect on the learning scores of subsequent A instances. Experiment 3 also provided examples of both negative and positive transfer. Adding B-instances to the pretraining block improved later A learning, a positive effect. But if we compare the pattern of results in Experiment 3 to those in Experiment 2, it appears that adding a single A-instance prior to even a fairly long series of B-instances exerted significant interference on later learning of both Bs and As. Thus, learning in the N=12 condition of Experiment 3 appeared poorer that that of the Contrast condition in Experiment 2, even though a longer block of same-category instances was presented in the former experiment. The difference was the single instance of the contrast category located at the beginning of the series in Experiment 3.

We argue that simple feature-associator models are unable to accommodate these seemingly contradictory effects, and a model with an explicit schema-triggering process is required. Simple associative models require simple effects; for example, increasing the number of instances from a given category should increase (or at least not decrease) learning of that category, and transfer effects between categories should be consistently positive or negative. Within the schema model, by contrast, our sequencing manipulations and their sometimes contradictory effects are readily interpreted as simple manipulations of triggering probability. Thus, the triggering model can accommodate the pattern of results obtained in these experiments whereas the simple associative model is strongly discredited by these results.

One question that has received little discussion so far is how schema-triggering explains the learning that did occur from mixed training sequences in these experiments. If long series of instances from the same category are required for schema-triggering to occur, then how does the theory explain learning in unblocked conditions where such sequences do not occur? One possibility is that categories are distinguished by schema-triggering when conditions are favorable for such triggering to occur, but that simple learning of inter-feature associations is the mechanism for learning when conditions are unfavorable for such triggering. According to this hypothesis, subjects do accumulate information about inter-feature correlations as they learn successive training instances, so that correlational patterns will eventually be learned regardless of sequence. However, this process would be slow relative to the much faster learning that occurs when subjects can explicitly separate the categories from the start and learn the defaults of each without interference from the other. This solution to the problem of learning from mixed sequences argues that there may be "strong" and "weak" forces in unsupervised learning, with the strong force (triggering of discrete schemas) producing more rapid learning and requiring less information be maintained in memory (it only requires tracking frequencies of individual values, rather than co-occurrence frequencies for all possible pairs of values), but also requiring fairly specific conditions for its occurrence.

A second hypothesis assumes that triggering of novel schemas can occur probabilistically in a mixed training sequence, but that the probability of this occurring at any particular point in the sequence is relatively low. According to this hypothesis, all category discrimination occurs as a discrete, all-or-none event (triggering), followed by a process of adjusting continuous strength values within each category. When a blocked sequence is employed, most subjects can be induced to trigger new schemas at the same point in training, e.g., at the first B-instance presented after a pure block of A-instances. When a mixed sequence is employed, triggering may occur at different points in training for different subjects. Some subjects might discriminate between categories virtually from the start of training, others might do so later in the sequence, and still others might fail to do so by the end of a given training session. The data from such a process, averaged over a group of subjects, would show much the same pattern of apparently gradual learning predicted by a single-process associative theory.

If this position is to be plausible, it must be shown that the conditions for probabilistic schema-triggering could occur within a mixed sequence. Imagine for a moment that the triggering criterion (the amount of mismatch between an instance and its source category required for triggering a new category) varies from subject to subject and from trial to trial within a given task. Given such variability, triggering might sometimes occur due to only mild discrepancies between an instance and its source category, i.e., when the instance violates relatively weak "defaults". Moreover, the strength of specific values in the aggregate schema would vary from trial to trial. For example, the strength of the A-default values could be temporarily increased in a mixed sequence whenever two or three A-instances appear in a row. If a B-instance were then encountered, its features might appear sufficiently surprising at that point to cause some subjects to create a new schema to describe that instance. In addition, the number of A-instances required for this would depend on how much impact each instance had on the norms of the schema, which could vary with subjects' momentary attentiveness and overall memorization ability.

The point is that schema-triggering might occur after initial aggregation if random circumstances were momentarily favorable, i.e., the values of one category were dominant in the aggregated norms because several instances of that category had occurred in sequence previously, a particular subject happened form strong encodings of these instances, and that subject's triggering criterion is momentarily lenient. If an instance of the contrast category occurs in such circumstances, it could result in the formation of a new schema. This is in contrast to the blocked conditions used above, in which circumstances highly favorable to triggering occurred at a single point in the training sequence, so that most subjects were induced create a new category at this specific point.

The results of these experiments provide strong evidence for the existence of a schema-triggering process, since this process is needed to explain learning in the blocked conditions. However, the present results do not strongly discriminate between these two explanations of unsupervised learning in mixed sequences. While parsimony favors the simpler explanation of no explicit recording of feature correlations (i.e., that all discrimination between categories is due to an all-or-none triggering event), these results do not rule out the possibility that learning inter-feature associations may also play a role in unsupervised learning.

Generality of the Results

One objection to generalizing from these results to unsupervised learning in the real world is that the stimulus variation in these experiments was rather artificial and stereotyped compared to the rich, complex variation typical of real-world domains. Of course, this point applies to most laboratory research on category learning, which commonly employs artificial stimulus sets generated from combinations of only three or four binary attributes. The purpose of the present experiments was to evaluate the attribute listing task as an index of unsupervised learning in a relativel, simple situation, and to use it to make elementary discriminations among models of learning in that situation. Demonstrating that a process such as schema triggering occurs under "artificial" conditions constitutes a prefectly valid proof of the existence of that process; it merely leaves the issue of boundary conditions unexplored.

Nevertheless, the basic attribute listing method could be used with many types of stimuli, including stimuli more complex and naturalistic than those used in the present experiments. For example, the present stimuli could be modified by adding several levels of continuous variation such as size or shade to each discrete value of each attribute. Thus, a set of insect stimuli could have two different styles of wings and three different size levels within each wing style. Such modifications would increase the background variability of the stimuli and could be used to make them appear more naturalistic. However, it should not change the basic pattern of results in the attribute listing task, i.e., a shift away from listing predictable aspects of the stimuli and an increasing focus on unpredictable information as the categories are learned.

Another sense in which the present stimuli appeared artificial was in the fact that the default values of each category occurred 100 percent reliability, i.e., attributes were perfectly correlated. This raises the question of whether anything about the current approach implies that categories must be defined by a set of necessary and sufficient features, an assumption that has been strongly criticized in recent years (e.g., Wittgenstein, 1953; Rosch, 1975, 1977; Smith and Medin, 1981). The schema-triggering model presented here makes no such assumption, although it does (reasonably) assume that people try to avoid violations of default expectations by forming subcategories to capture new patterns whenever they can do so. The model also assumes that since default violations are unexpected, subjects will tend to consider them highly informative and assign them a high attentional priority. This reflects the principle that violations of one's general beliefs are likely to be important both for distinguishing the particular situation in which the violation occurred, and possibly for making modifications in the general norms themselves.

Admittedly, the present experiments do not attempt to demonstrate unsupervised learning of categories with probabilistic features, which may limit the generality of the present results. However, the attribute listing procedure should be generalizable to learning problems in which category defaults are unreliable, assuming that people can learn such categories without feedback. It is clear from people's performance in the present tasks that many of the "fuzzy" categories used in standard supervised learning experiments, in which diagnostic features are highly unreliable would probably be unlearnable for most subjects without explicit feedback. This simply reflects the greater difficulty of the unsupervised learning task, in which subjects must generate their own categories and internal feedback. Interestingly, the schema-triggering model predicts that the effects of default reliability should depend on how the instances are sequenced. If subjects are shown "prototypical" training instances (in which all the expected default values are present) until they have learned to distinguish the categories (i.e., instances are assimilated to separate schemas), then the learning should be fairly resistant to later, non-prototypical instances in which individual defaults are violated. However, such instances could have a strong effect on learning if presented early in training, by interfering with initial discrimination between the categories. The investigation of such factors should provide interesting topics for future research, and allow many productive comparisons to be made between supervised and unsupervised learning.

References

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

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

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

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.

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.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93,* 136-153.

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.

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., & 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.

Rumelhart, D. E., & Zipser, D. (1986). Feature discovery in competitive learning. In J. L. McClleland, & D. E. Rumelhart (Eds.), *Parallel distributed process: Explorations in the microstructure of cognition, vol. 2.* Cambridge, Mass.: The MIT Press.

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.

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

Footnotes

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

[1] Five subjects were excluded from the data analysis, two from the Practice condition and three from the Contrast condition. These subjects were excluded because they produced no usable data from more than one third of the thirty two trials in the experiment. A subject was considered to have produced no usable data from a given trial if they listed no features on that trial (i.e., they left that page in the booklet blank), if the only information provided was a comparison to a previous instance (e.g., "same as the first one"), or if none of the features listed were representable within our eight attribute coding scheme.

Table 1

Design and Counterbalancing, Experiment 1

Attribute	Category(1)		Control(1)		Category(2)		Contrcl(2)	
1 (Wings)	1	2	1	2	1	2	1	2
2 (Body)	1	2	1	2	1	2	1	2
3 (Markings)	1	2	1-2	1-2	1-2	3-4	1-2	3-4
4 (Tails)	1	2	1-2	1-2	1-2	3-4	1-2	3-4
5 (Eyes)	1	2	1-2	1-2	1-2	3-4	1-2	3-4
6 (Legs)	1-2	3-4	1-2	3-4	1	2	1-2	1-2
7 (Jaws)	1-2	3-4	1-2	3-4	1	2	1-2	1-2
8 (Antennae)	1-2	3-4	1-2	3-4	1	2	1-2	1-2

Table 2

Design and Counterbalancing, Experiments 2 & 3

Attribute	Group 1		Group 2	
1 (Wings)	1	2	1	2
2 (Body)	1	2	1	2
3 (Markings)	1	2	1-2	3-4
4 (Tails)	1	2	1-2	3-4
5 (Eyes)	1	2	1-2	3-4
6 (Legs)	1-2	3-4	1	2
7 (Jaws)	1-2	3-4	1	2
8 (Antennae)	1-2	3-4	1	2

Figure Captions

Figure 1. Sample stimuli from Experiment 1. Instances of one category are on the right and instances of the other are on the left. The correlated attributes in this stimulus set are wings, abdomen shape, abdomen shading, mandibles, and antennae; the variable attributes are legs, tails, and eyes.

Figure 2. Attribute listing data plotted plotted over instances for the Control condition of Experiment 1. The two "pseudo-categories" are separated in these plots, but were presented in random order in the actual task.

Figure 3. Attribute listing data for the Blocked condition of Experiment 1.

Figure 4. Attribute listing data for the Mixed condition of Experiment 1. The categories are separated in these plots, but were presented in random order to the subjects.

Figure 5. Attribute listing data for the Practice condition of Experiment 2. The data for the two categories is presented separately and instances presented during the pretraining block are labeled "Pre-A" and "Pre-B".

Figure 6. Attribute listing data for the Contrast condition of Experiment 2.

Figure 7. Attribute listings for the three conditions of Experiment 3. Only the difference scores (listing of variables minus that of defaults) is shown here.

Figure 1.

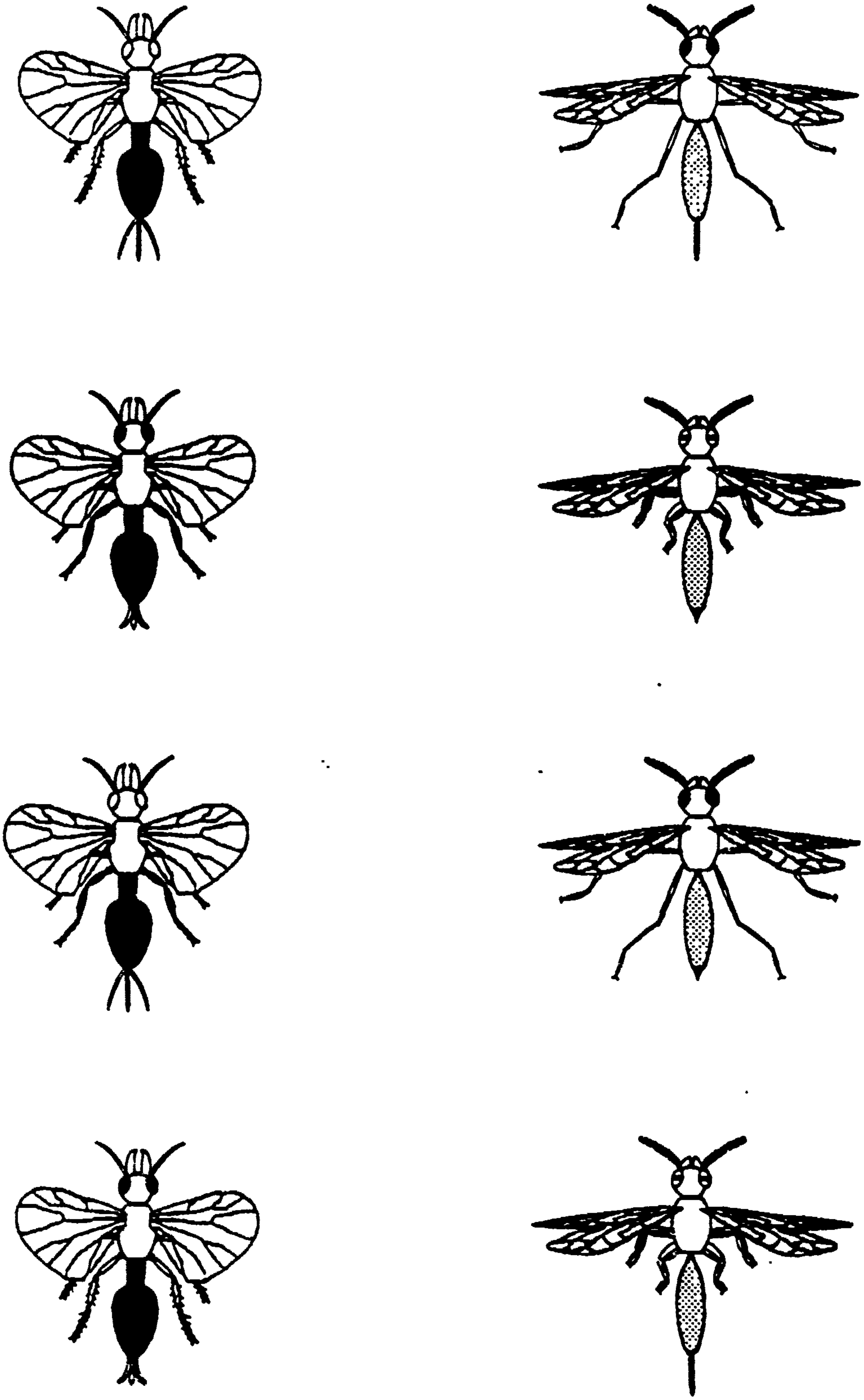
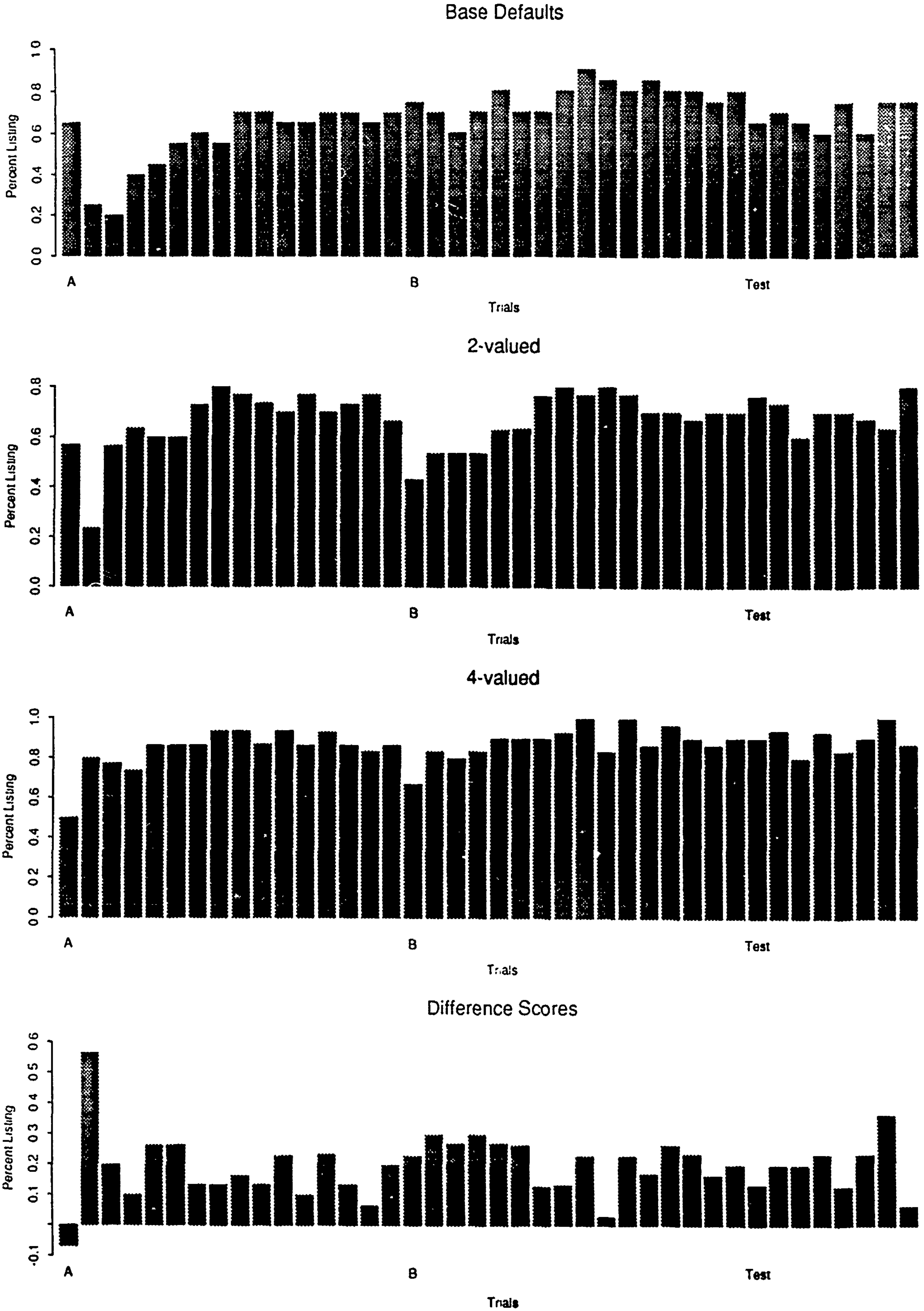


Figure 2



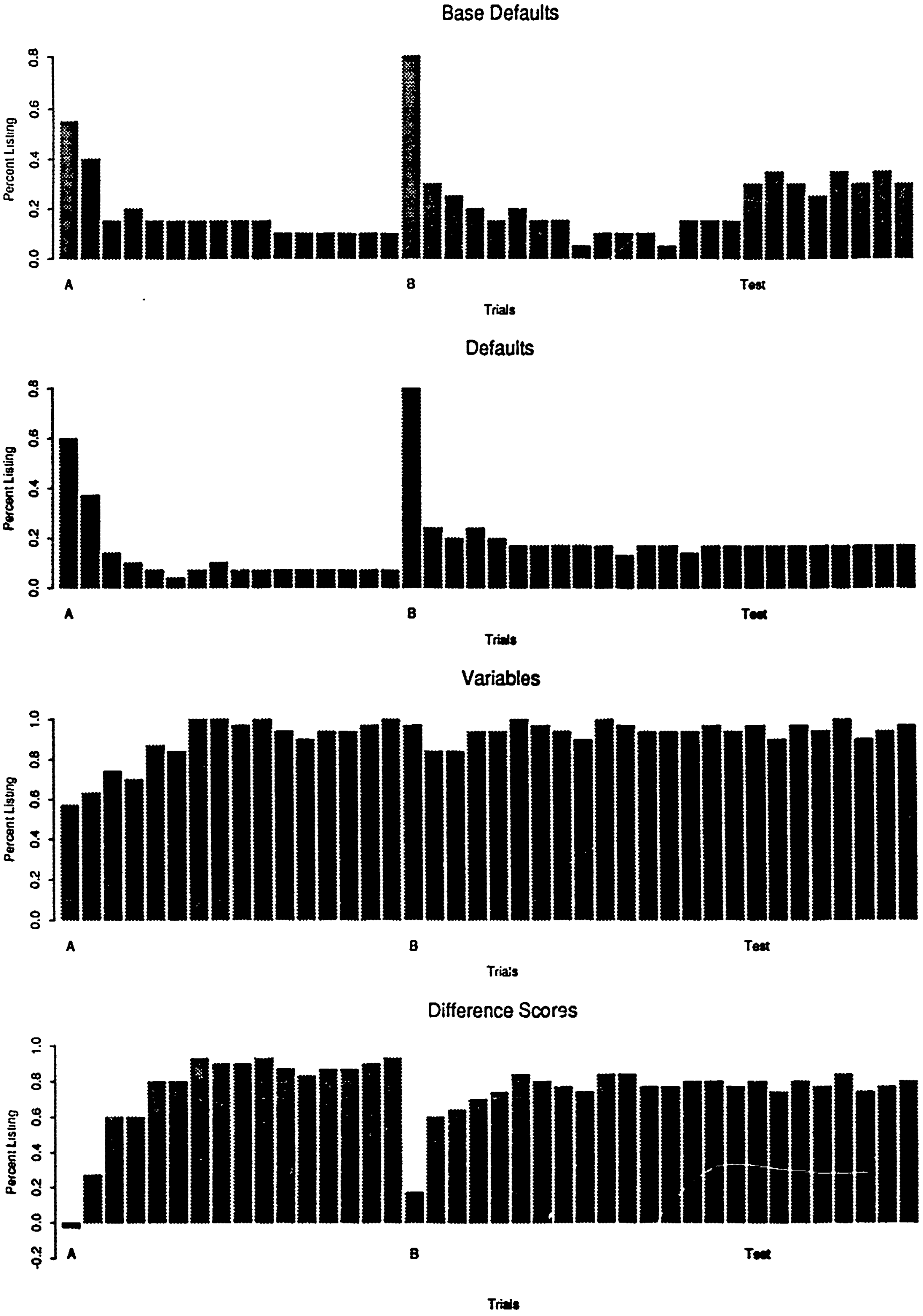
Base Defaults

2-valued

4-valued

Difference Scores

Figure 3



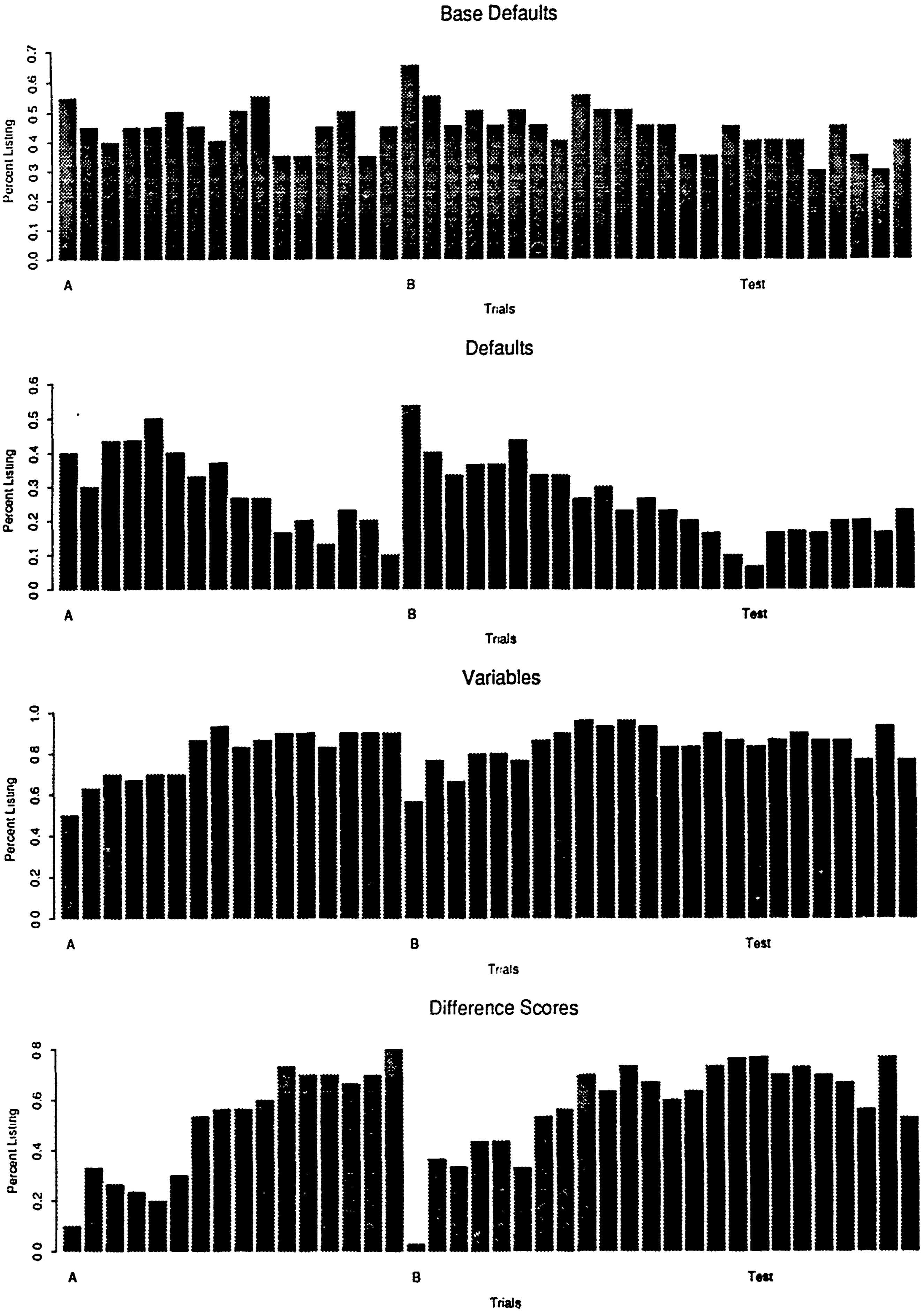
Base Defaults

Defaults

Variables

Difference Scores

FIgure 4



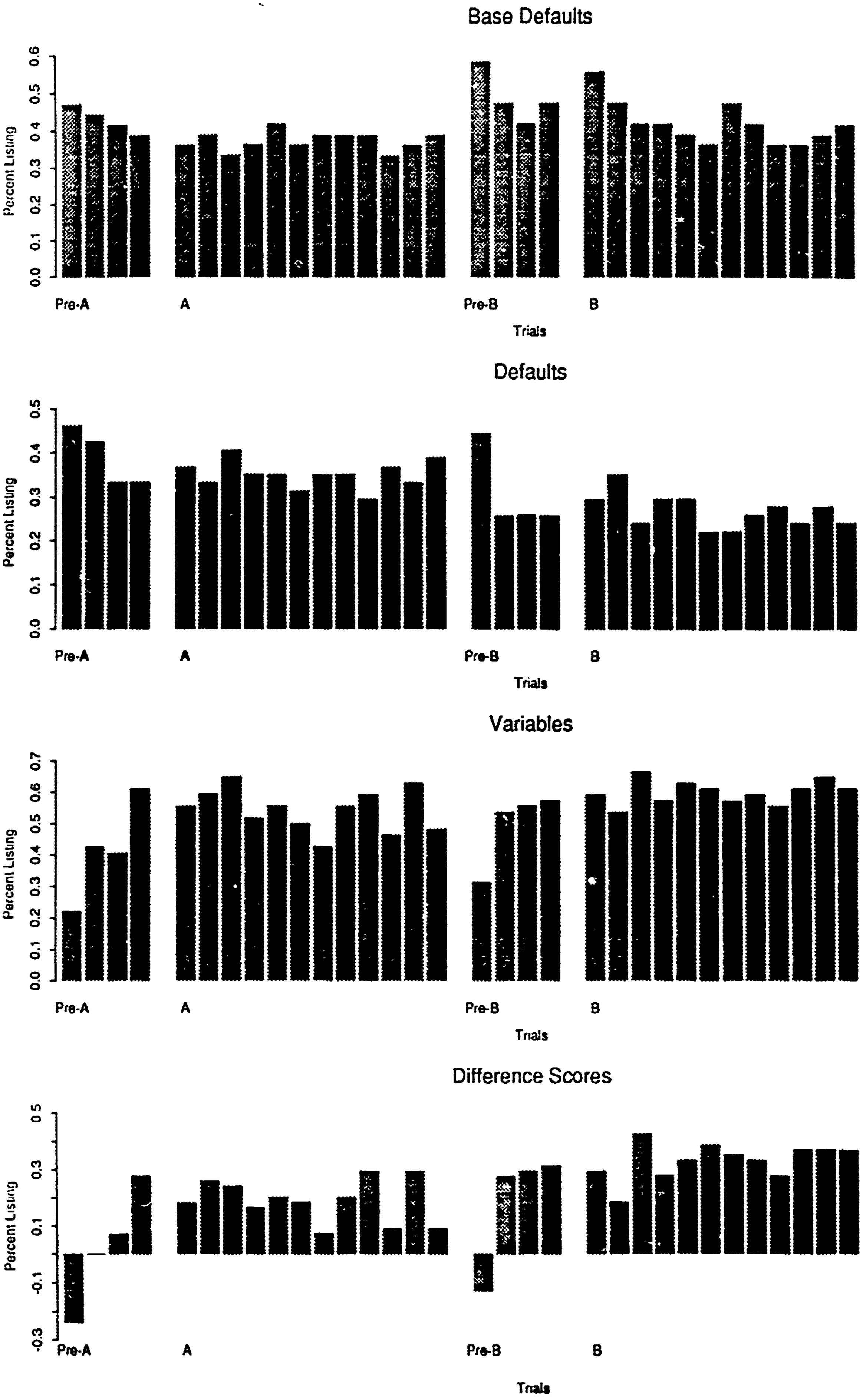
Base Defaults

Defaults

Variables

Difference Scores

Figure 5



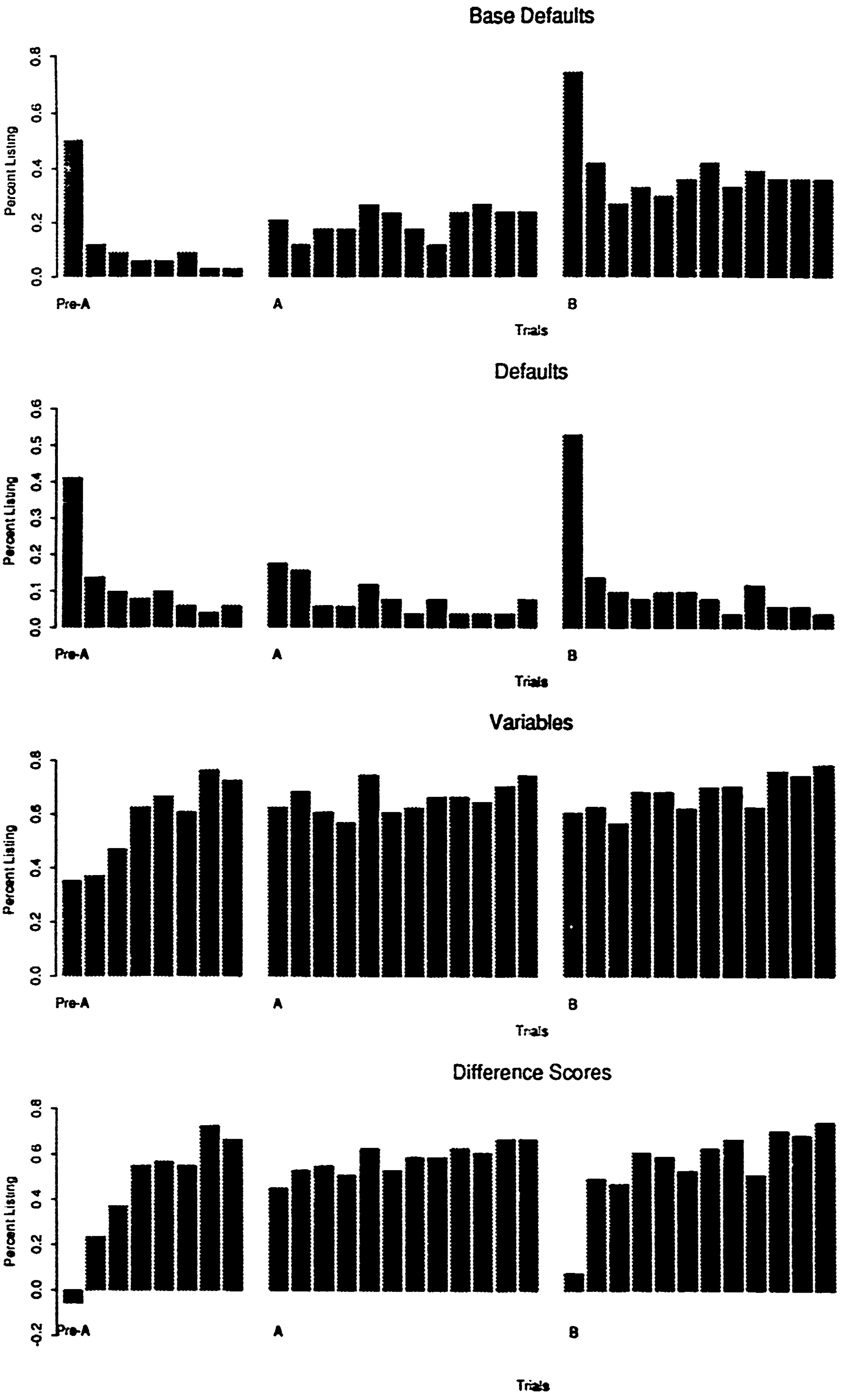
Base Defaults

Defaults

Variables

Difference Scores

Figure 6



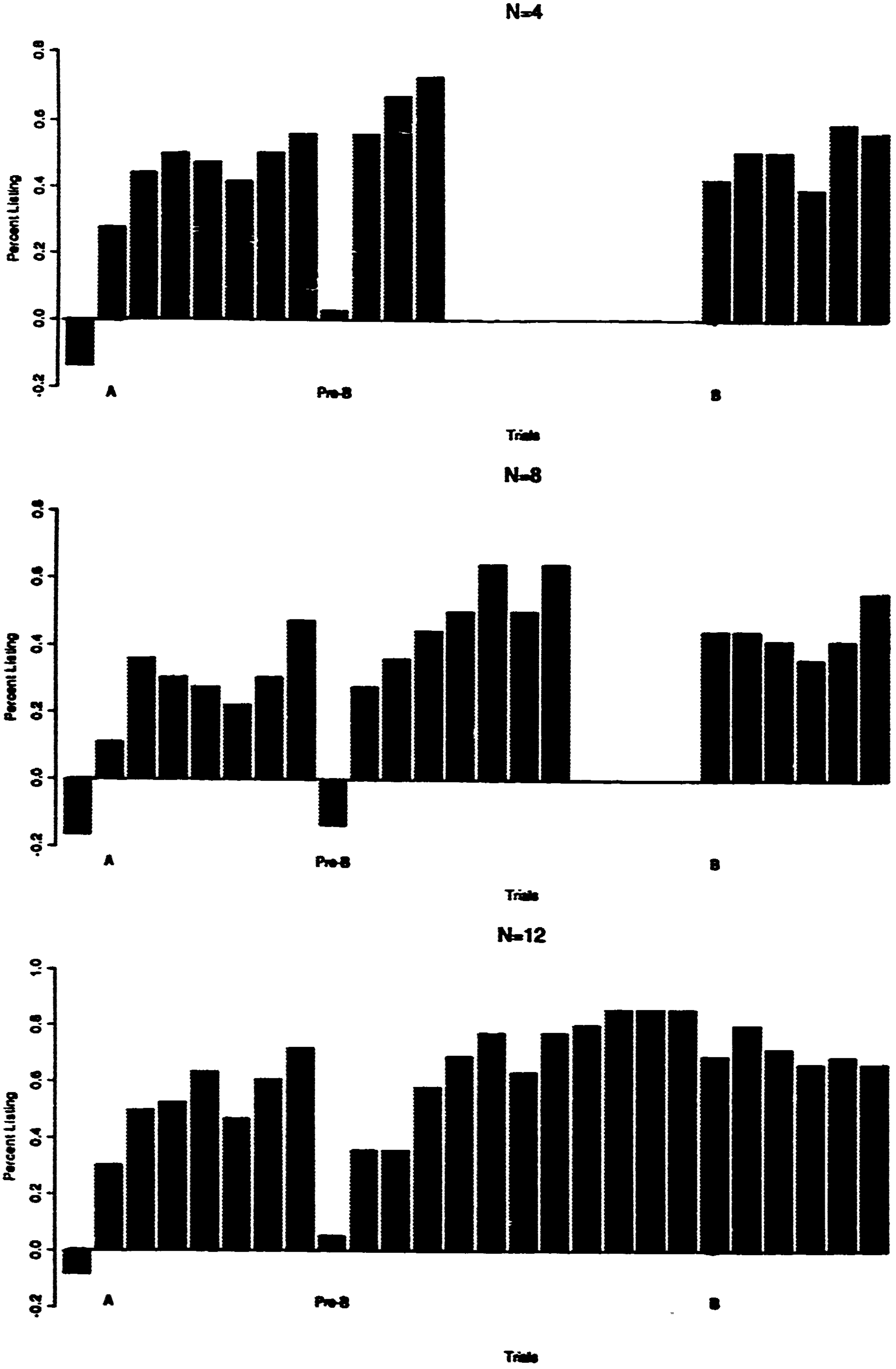
Base Defaults

Defaults

Variables

Difference Scores

Figure 7


N=4
Percent Listing
-0.2
0.0
0.2
0.4
0.6
0.8
A
Pre-B
B
Trials
N=8
Percent Listing
-0.2
0.0
0.2
0.4
0.6
0.8
A
Pre-B
B
Trials
N=12
Percent Listing
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
A
Pre-B
B
Trials