

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION

Unclassified

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

E

3. DISTRIBUTION/AVAILABILITY OF REPORT

Approved for public release;
distribution unlimited.

AD-A230 415

(S)

5. MONITORING ORGANIZATION REPORT NUMBER(S)

AFOSR-TR- 90 1175

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
ORGANIZATIONAir Force Office of
Scientific Research8b. OFFICE SYMBOL
(if applicable)

NL

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER

AFOSR-87-0282

8c. ADDRESS (City, State, and ZIP Code)

Building 410
Bolling AFB
DC 20332-6448

10. SOURCE OF FUNDING NUMBERS

PROGRAM
ELEMENT NO.
61102FPROJECT
NO.
2313TASK
NO.
A4WORK UNIT
ACCESSION NO.

11. TITLE (Include Security Classification)

Acquiring Generalizations to Organize Human Databases

12. PERSONAL AUTHOR(S)

Gordon H. Bower, John P. Clapper

13a. TYPE OF REPORT

Final Technical

13b. TIME COVERED

FROM 9/1/87 TO 8/31/90

14. DATE OF REPORT (Year, Month, Day)

1990, November 30

15. PAGE COUNT

17

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)
attention, concept, category, unsupervised
learning, default hierarchy

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

This report describes a three-year program of research on category learning in unsupervised environments, and the role of learned categories in the processing and retention of individual instances. A computational model of unsupervised category learning is described, and the model's implications for the evaluation, comparison, and memorization of instances are explored in several experiments. We introduce a new index of unsupervised learning, referred to as *attribute listing*, and show that such learning tends to optimize the encoding of instance features and their organization in memory. The empirical techniques developed in this project appear to hold considerable promise for further research on conceptual knowledge and its role in cognitive performance.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT

☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION

Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL

Alfred R. Freely, Ph.D.

22b. TELEPHONE (Include Area Code)

(202) 767-3021

22c. OFFICE SYMBOL

AFOSR/NL

ACQUIRING GENERALIZATIONS TO ORGANIZE HUMAN DATABASES

In order to act intelligently within their environment, biological or mechanical agents must possess an internal model of that environment and how their actions will modify it (see Craik, 1943; Johnson-Laird, 1983; Gentner and Stevens, 1983). The objective of this research was to investigate how humans learn internal models ("concepts") to characterize general categories of training instances (i.e., objects, events, or situations), and how these models facilitate the acquisition, organization, and retrieval of new information. We are primarily concerned with concept learning as it occurs in *unsupervised* environments, in which the learner must explore a domain of objects for themselves without a teacher, searching for regularities, consistencies, and clusters of correlated features in the objects. We assume that people use these regularities to invent or create subjective categories by which to organize the domain of objects, to control their expectations about and attention to these objects, and to guide their manner of recording specific instances into memory. Although unsupervised learning occurs continually in everyday life and in scientific discovery (where new concepts must be invented to deal with a novel domain), it has been little investigated in the psychological laboratory. These issues are probably central to a scientific understanding of human intelligence, since concepts are crucial to our abilities to learn, reason, and communicate. Furthermore, the research described here may shed light on central functional properties of human learners that could have direct practical application, by aiding in the design of training programs, instructional materials, computer-based learning systems, and the testing and selection of personnel.

- A. Clapper, J. P., Rehder, B., & Bower, G. H. A computational model of unsupervised concept learning. In preparation.

We have developed a computational model describing how people learn concepts and use them to guide their processing of specific instances during unsupervised learning. We view this model as instantiating a collection of heuristic principles that guide peoples' construction of category models as they explore a new domain. One of our assumptions is that people will create a new category in reaction to the failure of old ones to adequately describe an unusual stimulus. In so doing, they follow a heuristic principle, namely, if several properties of a new instance are surprising or inconsistent with one's norms for its assigned category, then a new category should be created to describe this unusual case. Within a category, we assume that people learn to ignore properties that consistently recur across instances and to selectively encode properties that cannot be predicted on the basis of their current model of the category. By this selection, learners reduce the amount of information they need to record to learn specific instances; the selection also causes learners to notice facts that might provide a basis for modifying or augmenting their current concept models. An informal description of the process model is provided below. The information processing in this model involves the following steps:

For	
HI	☒
ed	☐
tion	
on/	
Availability Codes	
Dist	Avail and/or Special
A-1	

1. *Retrieve the Best-Fitting Category.* When a novel stimulus is encountered, it is automatically categorized by matching a sample of its features (specific values of attributes) to a set of attribute-norms for each candidate category, and then selecting the best match. Examples of attribute-values would be large or small wings on an insect, long or short snout, rounded or angular head, and so on. The norms for a category are represented as a collection of *strengths* of association between the category and each value of an attribute. These strengths reflect the relative expectedness or availability of this attribute value, i.e., its frequency and recency among previous category members.

2. *Evaluate the Instance.* Once a stimulus is categorized, the norms for that category are used to interpret the instance and determine which features are most "informative" for learning about it. These features will then receive more attention or encoding resources. Although several definitions of "informativeness" are plausible, all capture the intuition that a value's informativeness increases with its unpredictability or surprise value. Importantly, this principle implies that consistent, highly expected values of an attribute (henceforth referred to as *defaults*) will be considered uninformative, whereas the informative features will be those that are unusual or not specified in advance by the concept. We are currently exploring several alternative indices of informativeness in our computer simulations, and will compare our simulation results to data from the proposed experiments to choose the best-fitting index.

3. *Encode the Instance.* After categorizing the instance and assigning informativeness to each of its values, the next step is to record the instance into memory. Here, we assume that the features of an instance compete for a fixed attentional or encoding capacity, which must be distributed among them in such manner as to maximize the informativeness of the features encoded. The model assumes that the amount of capacity allocated to encoding a given value is proportional to its informativeness relative to that of the other features of the stimulus. The encoding process produces a list of features with strengths stored in memory as the persisting trace of the instance. A feature's strength in this record depends on how much attention it received at encoding, which depended in turn on its informativeness.

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

(A) *Assimilation to a Previous Category.* Normally, instances are assimilated to the category used to evaluate and encode them. The norms of this category are adjusted by increasing the strength of each presented value in proportion to how much attention it received during encoding. One consequence of this updating rule is that repetition of a value has progressively diminishing effects on its strength in subjects' category norms. That is, because subjects pay relatively little attention to default values, new instances cause little change in their existing strengths.

(B) *Create a New Category before Assimilating.* New concepts are triggered by surprise, i.e., whenever multiple failures of the subject's category expectations (two or more in the simulation) occur together in the same instance. (An "expectation failure" is

defined as any value whose informativeness exceeds some internal criterion, and so is taken as inconsistent with the subject's norms for that attribute). If a new category is triggered by an unusual instance, then the instance will be assimilated to it and will not affect norms for previously-existing concepts.

The new concept is created by (1) generating new norms for the unusual attributes which caused the expectation failures, and (2) transferring norms for the remaining attributes from the "source" concept originally used to interpret the instance. To create a new norm for an attribute, the model assumes that subjects "reset" the strength of all its values to a low, baseline level, and then increase the strength of the presented value in proportion to the large amount of attention it receives during encoding. We assume that the source and new category share norms for attributes that have the same default values; these shared norms characterize a more inclusive (superordinate) class that includes both the new and source concepts as subordinates. In this manner, continuing exploration of a domain tends to build up a nested default hierarchy based on superordinate and subordinate relations and property inheritance among the concepts.

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

 Insert Figure 1 about here

Our assumptions about encoding and retrieval imply differences in the way default versus distinctive features of an instance are remembered. Because of their predictability, the default features of a category should have very low strengths of association with particular instances. To a first approximation, we may assume that subjects omit these features entirely from their memory representations of specific instances. Rather, the defaults are noted as properties of the general category and these can be inferred for specific instances by property inheritance. In such a memory organization, category defaults are effectively segregated off from the distinctive features of the individual instances (see Figure 1). The exemplar with its distinctive features is recorded as a "subnode" in memory pointing to the category node with its associated defaults. As a result, in retrieving instance-distinctive features, the system avoids fan effects due to category features. This "subnodding" confers a major advantage on this memory organization for later information retrieval, in addition to the economy of learning and storage that results from the encoding process used to create it. This

memory organization helps solve the so-called "*paradox of interference*", which is that experts with vast domain knowledge do not have the slowing of retrievals that interference theory alone would have expected (Smith, Adams, & Schorr, 1978). Our subnodding solution is similar to earlier solutions of the paradox that were proposed by Reder & Ross (1983) and Anderson (1983). We will return later to this topic.

Comparison to Alternative Approaches

Our model differs in several ways from previous models of unsupervised learning. As one example, our model learns incrementally, modifying its category norms in response to each presented instance. This property contrasts to several statistical clustering models (e.g., Michalski & Stepp, 1983; Fried & Holyoak, 1984) that do not learn incrementally; i.e., those models do not alter norms from individual cases examined one by one, but instead compute parameters of a classification scheme after analyzing a complete set of instances. Incremental learning is an attractive property for a psychological model because humans are clearly capable of learning on a case-by-case basis; moreover, humans are sensitive to the order in which instances are seen, whereas the omniscient AI models typically arrive at concepts that are independent of the order of seeing examples. A second advantage of the present model is that it makes explicit the role of generic concepts in the interpretation, analysis and recording of novel cases; in turn, the model shows how the processing of specific instances affects the learning of category-level expectations. Most previous models of category formation are strictly "bottom-up", in the sense that they specify how instance information is used to form concepts but not how the concepts in turn determine the encoding and representation of further instances. By exploring these issues in theory-guided experiments, we hope to shed light on how concepts function in normal cognition, an issue that has not been emphasized by previous theories of category learning.

Importantly, most previous models of concept learning were formulated to deal only with the classification of instances into categories, and did not consider the problem of storing those instances in memory for later reproduction. Consequently, they assume that learners become more likely to attend to attributes whose values consistently co-occur across category members (i.e., that are diagnostic of category membership, e.g., Billman & Heit, 1988). While this process is acceptable for classification, it is not adequate for learning and retrieving descriptions of specific instances. For example, such increasingly focused sampling would lead to less and less learning about the distinguishing features of specific instances. In addition, a learning process that focuses solely on classification will be blocked or very slow in learning specific subcategories that are differentiated within more general categories. For example, once having learned to differentiate oak trees from maple trees, people operating under this limitation would be prevented from attending to more subtle properties that differentiate various subspecies of oaks because they would be focusing instead only on features common to all oaks. Such a focus contrasts with more naturalistic learning, in which people consider known categories as "background", and proceed to focus on more subtle distinctions between instances that might form a basis for learning more differentiated categories.

- B. Clapper, J. P. & Bower, G. H. The role of category knowledge in encoding and remembering instances. Submitted for publication.

The memory experiments described in this article were designed to demonstrate that people would learn and represent instances of a well-known category in terms of their general schema of that category, as predicted by our information-processing model.

Experiment 1

This experiment aimed to show that when learning instances of a well-known category, defaults and non-defaults would be stored separately in memory, and only non-defaults would be explicitly recorded as facts about individual instances. In this experiment, subjects learned instances and categories of astronomical stars, described by lists of constituent chemical elements. They were first several taught categories defined by collections of co-occurring elements; we then examined their learning and later retrieval of the properties of specifically-named stars (instances). Although we did not mention the parent category of an instance, each instance possessed all the default features of its parent category, plus one or more extra features not universally present in instances of that category. We found that instances with more of these distinctive features took subjects longer to learn, but that the number of category defaults possessed by an instance had no effect on subjects' learning rates. This indicates that subjects did not record such predictable defaults when learning the instances, consistent with our theory of a model-based encoding process. We also found that the time required by subjects to verify the features of an instance in a later recognition test increased with the number of distinctive features the instance possessed, but not with the number of category defaults it had. This lack of fan effects suggests that defaults were retrieved from the general schema rather than being stored directly with instances, as the distinctive features were. This segregation of general and specific information could function to prevent general knowledge that a person accumulates about a category from interfering with their ability to retrieve facts about specific instances.

Experiment 2

The results of the previous experiment were highly consistent with the model, but that experiment could be criticized on the grounds that subjects were directly taught the categories we wanted them to know, and so the results might not apply to situations in which categories are induced from individual instances. Thus, a second experiment was constructed to extend those earlier results to a task in which subjects induced concepts for themselves in an unsupervised environment (i.e., they were not given corrective feedback to help them discriminate instances of the different categories). In this experiment, subjects had to study then recall many stimuli (4- to 6-item sequences of uppercase consonants) in a long series of study-test cycles similar to the familiar "Brown-Peterson" short-term memory task. The stimuli were from two different categories, each defined by a different group of co-occurring properties (specific letters-in-positions); we expected that subjects would discover these categories from their experience with the training instances, and use them to improve their recall performance. As in the previous experiment, the results showed that, once having learned the regularities in the first few training instances, subjects' learning of further instances was

affected only by the number of distinctive features (letters) they possessed, and not by how many defaults they had. Their learning of the distinctive features of category members was also improved relative to the corresponding features of stimuli for which no generic concept had been learned, consistent with our hypothesis that a concept would help learners to focus their encoding capacity on such informative properties. This result is consistent with the attentional biases and performance benefits that we expect to accompany the learning of an internal model of a category.

Experiment 3

A factor that limits the generality of the preceding two experiments is that in both the defaults were present with 100 percent reliability in every exemplar of the experimental categories. By contrast, real-world categories are often somewhat indeterministic in the features their members possess, and even highly typical features (say, "flying" for the category "birds") may be absent or altered in specific instances (e.g., penguins are birds that swim but cannot fly). Thus, we designed a second short-term recall experiment, similar to the last except that only a single category was presented to each subject and the defaults (specific letters-in-positions) were occasionally missing and replaced by an alternative, surprising value (letter).

In this experiment, the reliability of the defaults (i.e., percentage of instances in which they were present) was varied across different groups of subjects (60, 70, 80 and 90 percent, plus a control group that saw randomly-constructed instances for which no defaults could be learned). We found that the recall of variable (unpredictable) features of the instances was higher the more predictable were the defaults for a given group. That is, the greater the proportion of instances in the group that possessed a given default, the less informative subjects considered it to be for recording any particular instance; thus, the default would compete less with the variable attributes for attentional resources at encoding. We also found that subjects within each group showed poorer recall of the variable features of an instance the greater the number of default exceptions it contained. Such exceptions possess a high level of discriminative informativeness and compete strongly with the other features of an instance for attentional resources during encoding, thereby reducing their initial learning and later recall.

Importantly, although decreasing the predictive reliability of subjects' category models reduced their functional benefits on recall performance, subjects were still able to exploit their models to some extent even when the defaults were fairly unreliable, i.e., recall benefits were observed when defaults were present in only 70 or 80 percent of the instances. Thus, this experiment shows that our theory of unsupervised learning generalizes to domains in which defaults occur probabilistically, and that the functional benefits that accompany learning of a category model depend on the overall reliability of the model's predictions and its degree of match to specific instances.

- C. Clapper, J. P. & Bower, G. H. The impact of category knowledge on the similarity of instances. Submitted for publication.

Experiment 1

Our model's prediction that people primarily attend to distinctive information while ignoring category defaults can be tested by investigating people's judgments of how *similar* various category members should appear to each other. We predict that a feature's weight or impact on a judgment of the similarity of two category members should be proportional to its informativeness, as specified by that category. Therefore, unexpected, novel, and surprising properties of the stimuli should tend to dominate subjects' comparisons, while predictable category features should tend to be ignored.

Our first experiment to test this hypothesis consisted of a series of similarity judgments (on a 20-point scale) between many pairs of instances of a single category (pictures of fictitious insects that varied in several attributes, see Figure 2). In some of these pairs, one or both of the instances were missing a feature that was normally present in members of that category. When this surprising absence occurred as a *common* feature of the instances, we expected their similarity to be increased relative to an otherwise-equivalent pair in which the default value was present. And when the instances differed on a normally-consistent attribute, e.g., when one instance had the default value and the other had a different, unexpected value, we expected that this surprising difference would reduce their rated similarity more than a comparable difference between two familiar values of a highly variable attribute. The latter prediction was confirmed by the data, but in this first experiment we found little evidence that pairs in which both instances lacked an expected default would be rated more similar than normal pairs.

Insert Figure 2 about here

Experiment 2

We ran several further experiments to pursue our "shared absence" hypothesis; unfortunately, subjects in these experiments tended to rate similarity by simply counting differences between the instances and ignoring their common features, so that our manipulations of common features had little effect on their ratings. We have now overcome this problem in new experiments in which subjects rate the similarity of specific pairs from memory, given only the names of the instances learned earlier. Since experiences tend to be remembered by their distinctive or unusual properties, we expected similarity ratings of remembered instances to be dominated by their shared exceptions. In a first experiment of this type, we obtained a significant increase in similarity due to the shared absence of a default. This result confirms our assumptions about the effects of category learning on the underlying memory representations formed of stimulus patterns, and indicates that this similarity rating task may have considerable promise in further research on these issues.

- D. Clapper, J. P. & Bower, G. H. Learning and applying category knowledge in unsupervised domains. To appear in *The Psychology of Learning and Motivation*, Vol. 27, G. H. Bower (Ed.), New York: Academic Press, 1991.

This chapter describes the experiments listed below, plus the information-processing model and several of the experiments described earlier.

Experiment 1

A distinctive feature of this project is our concern with observational or unsupervised learning, because of its importance in naturalistic learning and because characterizing the concepts that can be so learned might help clarify the vague but important notion of a "natural" concept. To investigate such observational learning, we have developed a new task that allows the evolution of a category model to be observed on a trial-by-trial as it is being learned. In this task, subjects are shown a series of instances and asked to list the distinguishing characteristics of each. In a first experiment of this type, subjects were asked to list the distinguishing features of a series of instances from a single category. After the first few trials, they mentioned the presence of expected (default) attributes much less frequently than variable attributes, indicating that they had learned that the presence of these features could be taken for granted (see Figure 3). Moreover, when a default was absent from a specific instance, subjects were highly likely to note this surprising absence; listing of this attribute would then be elevated for several trials before returning to baseline. The overall pattern of listings resembles the cycle of habituation to a repeated stimulus, dishabituation to an unexpected change in the stimulus, and gradual re-habituation over the following trials.

 Insert Figure 3 about here

Experiment 2

This experiment extends Experiment 1 by demonstrating the spontaneous acquisition of two contrasting categories in an incidental learning situation (i.e., subjects were not explicitly asked to search for categories or correlational rules among the instances). The categories were distinguished by different default values on several attributes. Instances of one category ("Category A") were presented for a first block of 16 trials, and then instances of another category ("Category B") were presented for a second block of the same length; a series of eight transfer trials then ensued in which instances of the two categories were randomly interspersed in the sequence. As expected, subjects gradually learned the Category A defaults during the first block of trials, and decreased their listing of these properties accordingly. When they encountered the first instance of Category B, which had contrasting values on several of the consistent attributes of Category A, subjects at first greatly increased their listing of these attributes. Over the next several trials, they gradually reduced their responding to these attributes and returned to listing only the variable features of instances as the default values of the new category were learned. These listing patterns reveal orderly learning curves for the acquisition of the two concepts (see Figure 4). Importantly, subjects did not show a

significant increase in their response to the the default values of either category during the mixed block. This indicates that they had acquired stable concepts which they could apply across different contexts, rather than merely showing local patterns of habituation and dishabituation caused by "runs" of instances with repeated values.

 Insert Figure 4 about here

Experiment 3

Our theory assumes that people create new category models mainly in response to failures of their previously-existing models. The stronger the specific expectations that are violated by a given instance, the greater should be the likelihood of creating a new category around that instance. In the previous experiment, we obtained clear learning of both A- and B-categories by presenting the instances blocked by category, so that the subjects had time to build up strong A-defaults prior to encountering their first B-instance. This training sequence maximized the probabilities that the contrasting B-defaults would appear highly surprising and trigger a new category, rather than merely being assimilated into the existing Category-A norms. By contrast, our theory expects that interspersing A- and B-instances in random sequence from the start would interfere with subjects' learning to discriminate the separate categories; this should occur because subjects would have seen only a few A-instances before the first B-instance was presented, increasing the chance that they would start off assimilating instances of the two categories together into a common set of norms.

We recently completed an experiment in which we tested this prediction. The stimuli were the same as those used in the previous study, except the instances of the two categories were presented in a randomly interspersed sequence rather than being blocked by category. The pattern of attribute listings from this experiment were consistent with our theory's predictions about sequencing effects. Significant learning of the category defaults was observed, i.e., subjects learned to list variable attributes significantly more often than defaults by the end of the experiment. However, it is clear from inspection of the data displayed in Figure 5 that this learning was much poorer than that observed in the prior blocked-by-category experiment. Subjects' listing of defaults in this experiment never declined to the low level that they did in the previous studies. Moreover, what learning did occur was accomplished much more slowly when the categories were interspersed than when they were separated in the training sequence. This strong interference with category differentiation due to interspersed presentation is predicted by our theory, but several other current models of unsupervised learning (especially those in which the system learns by sampling and testing specific hypothesis, e.g., Billman & Heit, 1988) cannot readily accomodate this finding.

 Insert Figure 5 about here

Experiment 4

In accumulating knowledge about a domain, people often develop a series of related categories at multiple levels of specificity (see, e.g., Holland, Holyoak, Nisbett, & Thagard, 1986). Many real-world domains, such as categories of animals, plants, automobiles, jet aircraft, and medical diseases, are partitioned at more than one level, as some form of default hierarchy. Despite the prevalence and importance of conceptual hierarchies, prior research on category learning has usually examined only single-level categories. The aim of this experiment was to test whether our attribute listing paradigm could be used to study category learning in hierarchically organized stimulus domains.

This experiment was similar to Experiment 2 except that four categories were used instead of two. Instances of the first category -- call it A1 -- were presented for the first ten trials, followed by ten A2-instances, then ten B1's, and ten B2's. The default values characterizing the four categories can be denoted as follows: A1 = 111111XX, A2 = 111222YY, B1 = 222333QQ, and B2 = 222444RR, where X, Y, Q, and R denote different pairs of values of variable attributes occurring in each of the four categories. As previously, subjects were asked to list the distinguishing features of each instance.

The results are displayed in Figure 6. Subjects decreased their listing of A-defaults (attributes one through three) throughout the A1-instances and showed no increase during A2, which shared these values (panel A of Figure 6). A marginally significant increase in listing these attributes occurred for the first B1-instance ($p < .10$), followed by a rapid decrease back to the zero baseline. For subordinate level defaults (attributes four through six), listings reflected similar patterns of co-occurrence and contrast, increasing sharply for each new subcategory and decreasing over successive instances of the same subcategory (panel B of Figure 6). Importantly, reporting of defaults at either level showed no significant increase on the first instance of the mixed block. During this block, listing of defaults was over 80 percent lower than listing of variable attributes (panel C), a highly significant difference.

Insert Figure 6 about here

These results indicate that subjects learned stable categories during the training blocks and applied these categories to interpret instances during the mixed test block. The patterns of listings also showed that subjects transferred superordinate defaults across subcategories, since there was no increase in listing superordinate A-defaults when the first instance of A2 was encountered, or of B-defaults for the first instance of B2, but increases did occur for the changed, subordinate, defaults. These results indicate that our subjects learned to distinguish multiple categories within the hierarchically organized domain. However, more research will be required to characterize details of how such knowledge is organized in subjects' memories, and to identify boundary conditions and major variables that influence learning in such domains.

References

- Anderson, J. R. (1976). *Language, memory, and thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, 22, 261-295.
- Billman, D., & Heit, E. (1988). Observational learning from internal feedback: A simulation of an adaptive learning method. *Cognitive Science*, 12, 587-625.
- Craik, K. (1943). *The nature of explanation*. Cambridge: Cambridge University Press.
- Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10, 234-257.
- Gentner, D., & Stevens, A. L. (1983). *Mental models*. Hillsdale, NJ: Erlbaum.
- Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning and discovery*. Cambridge, MA: MIT Press.
- Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge: Harvard University.
- Michalski, R. S., & Stepp, R. E. (1983). Learning from observation: Conceptual clustering. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach*. Palo Alto, CA: Tioga Publishing Company.
- Reder, L. M., & Ross, B. H. (1983). Integrated knowledge in different tasks: The role of retrieval strategy on fan effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9, 55-72.
- Smith, E. E., Adams, N., & Schorr, D. (1978). Fact retrieval and the paradox of interference. *Cognitive Psychology*, 10, 438-464.

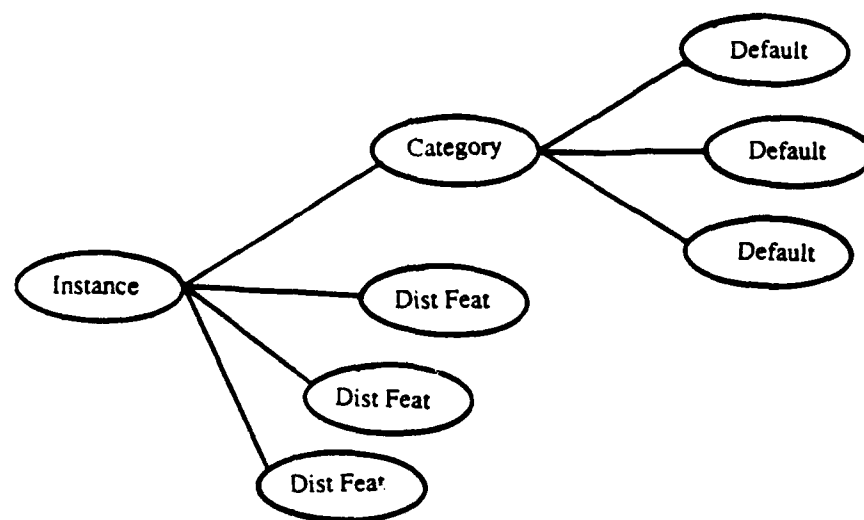


Figure 1. A network representation of an instance and its distinctive features encoded as a "subnode" of a general category. Each line in the figure represents an associative connection. Because defaults are stored with the category, increasing their number should produce no fan effects on retrieving instance-to-distinctive feature associations.

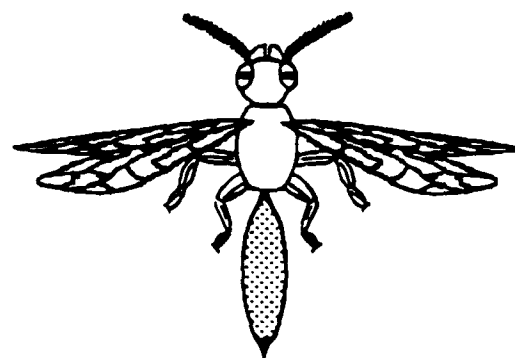
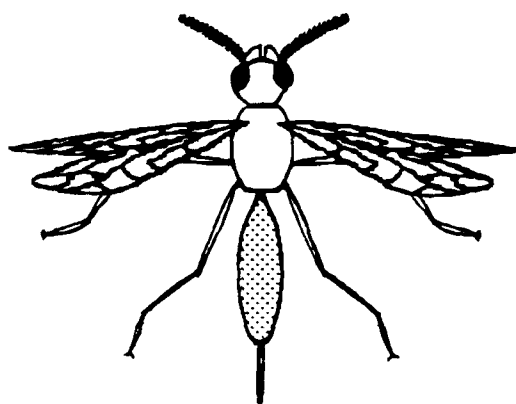


Figure 2. Sample stimuli from our similarity experiments. These fictitious insects share several default attribute values within a category, and differ along several variable attributes. The instances depicted above are all from a single category.

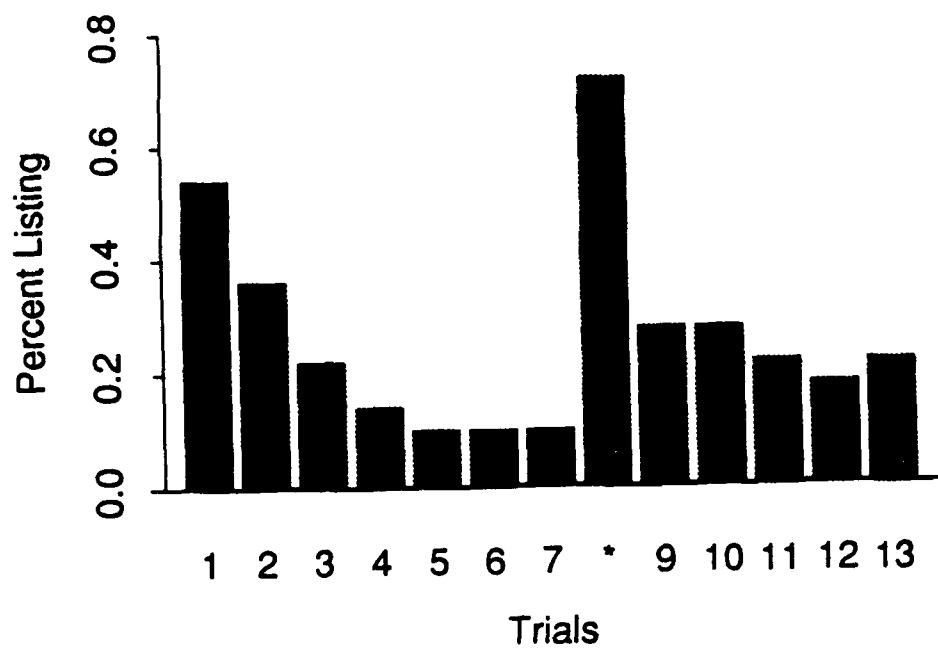


Figure 3. Observed percentage of default values listed by subjects over successive instances of a single category. An instance lacking the default was presented on trial "*"; the default was present in all subsequent instances.

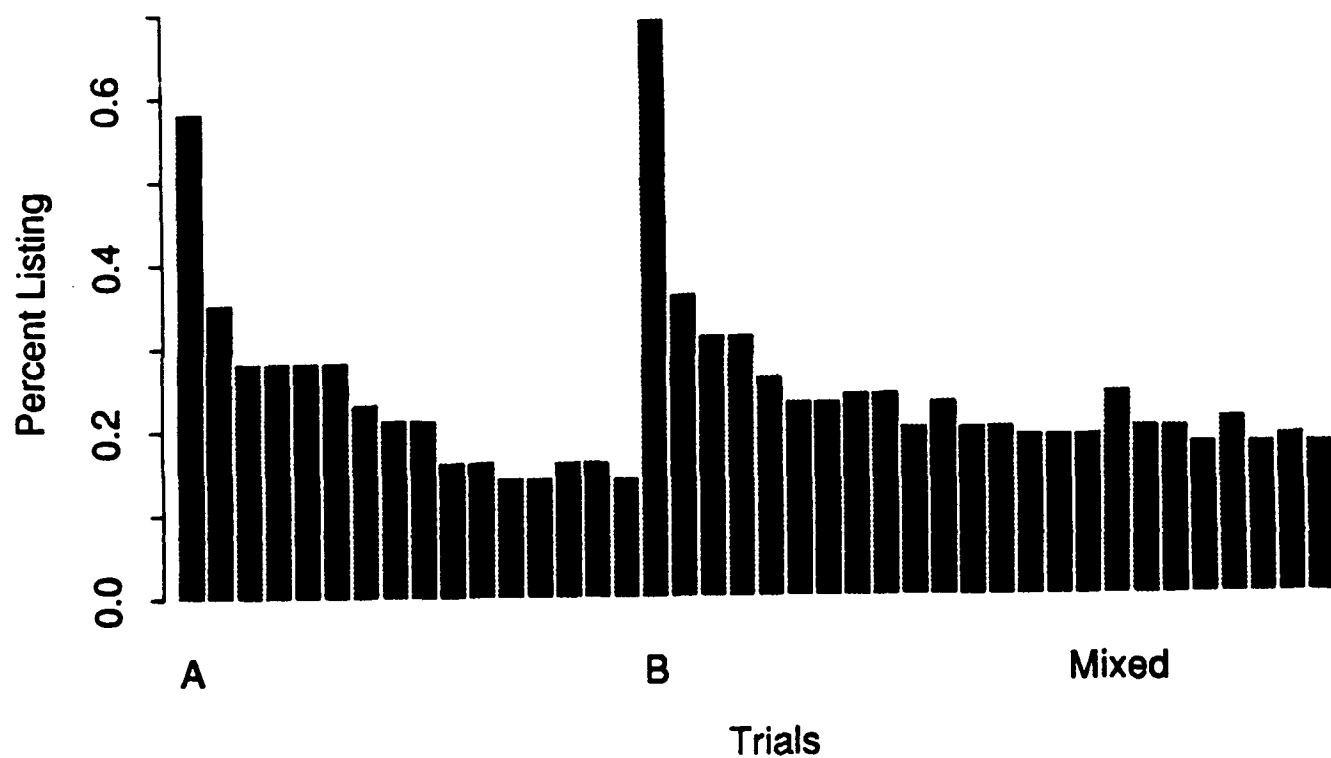


Figure 4. Observed percentage of default values listed by subjects in a two-category experiment. The default values were switched to Category-B at trial B. Instances of both categories were presented in random order during the final, mixed block.

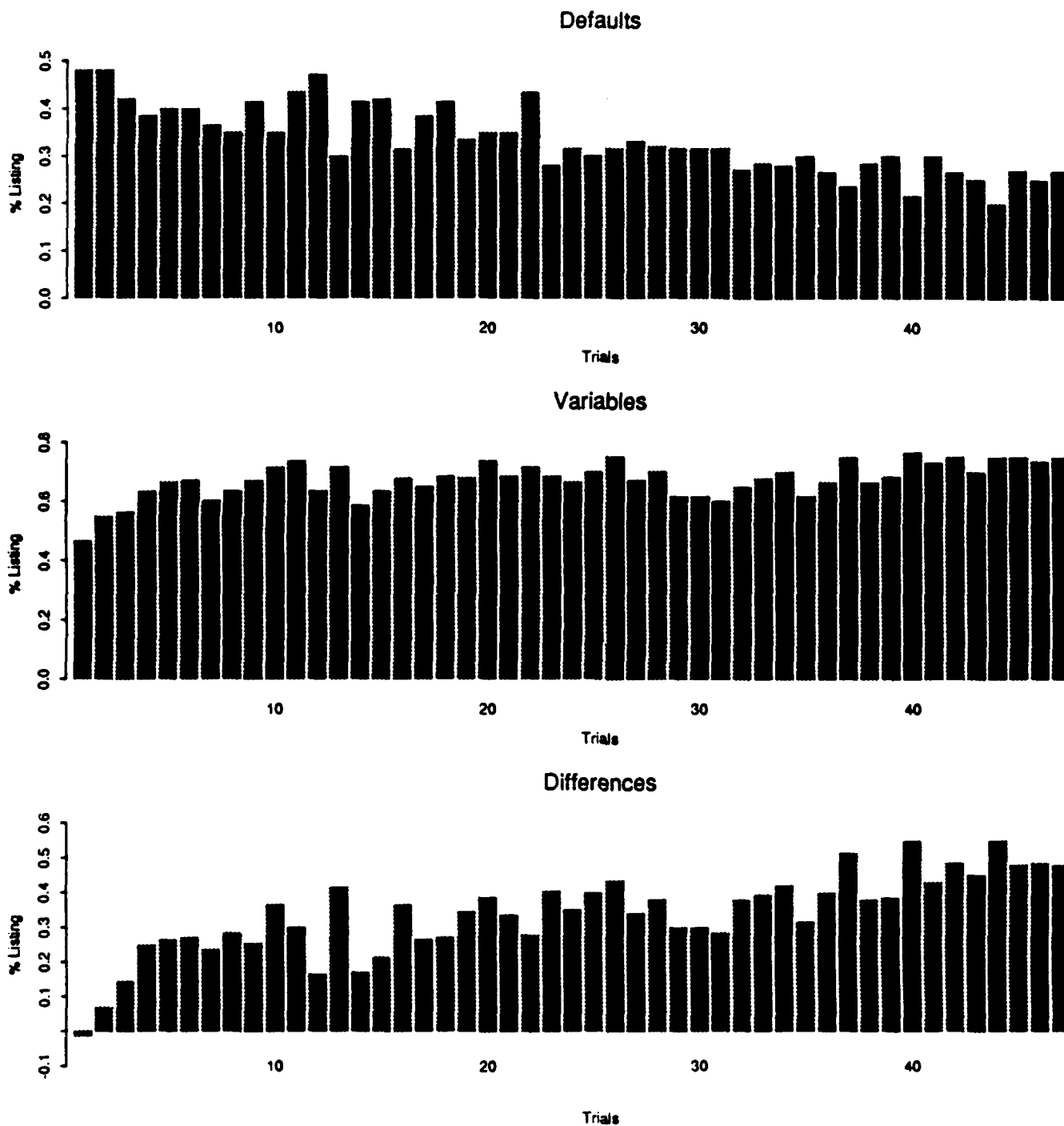
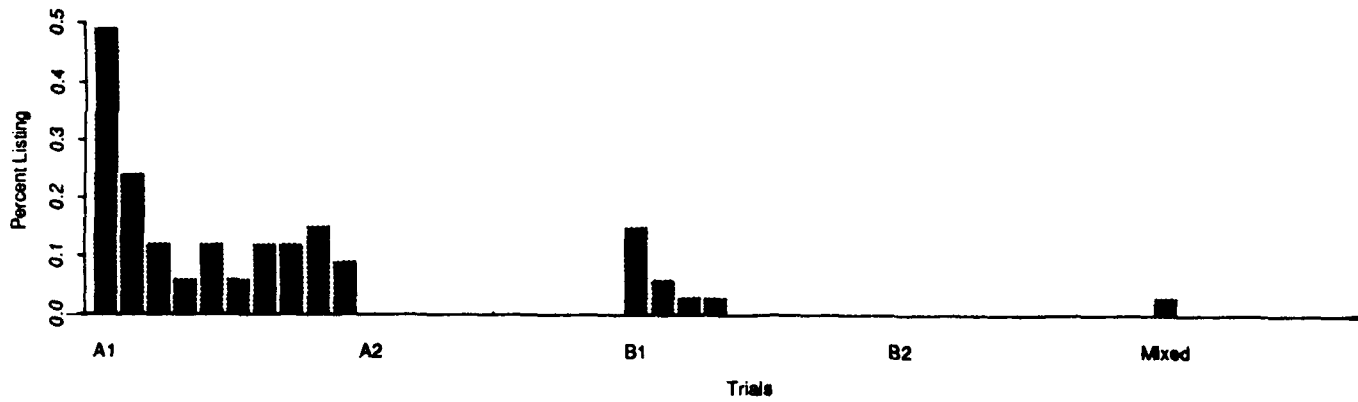
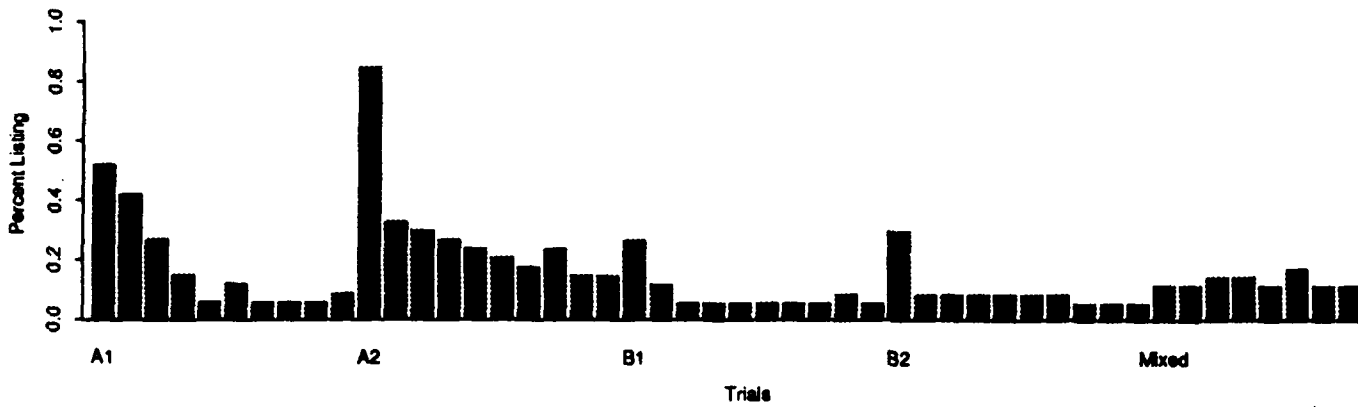


Figure 5. Observed percentage of defaults (top) and variables (middle) listed in an experiment in which two categories were presented in a mixed sequence. The bottom plot shows the difference in listing frequency between the two feature types.

Superordinate Defaults



Subordinate Defaults



Variables

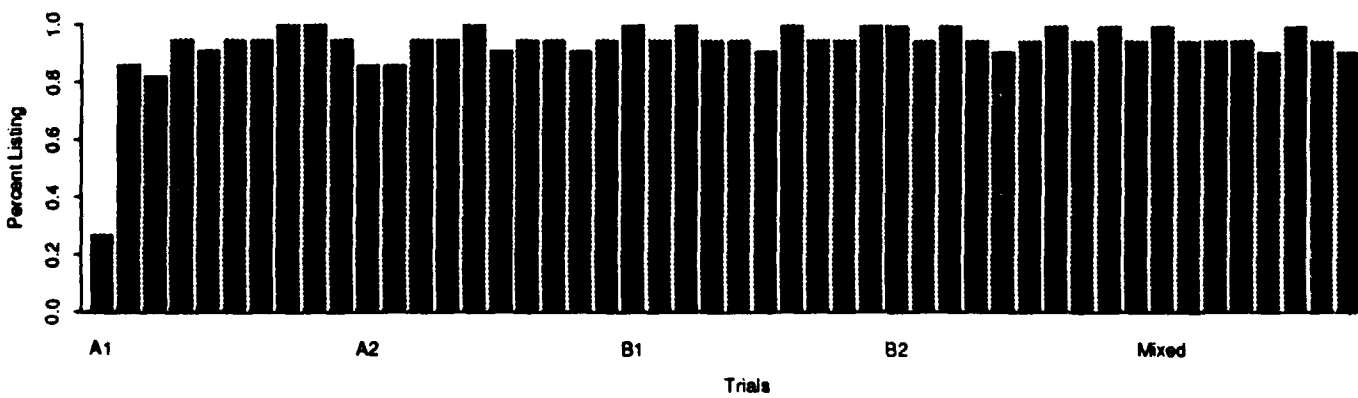


Figure 6. Percent listing of superordinate defaults (top), subordinate defaults (middle), and variables (bottom) from Experiment D4.