Learning Categories Without Teachers

By

Gordon H. Bower and John Clapper
Stanford University

Paper presented at "Practical Aspects of Memory" Conference. University of Maryland. August 2, 1994

Research Supported by the Air Force Office of Scientific Research, Grant

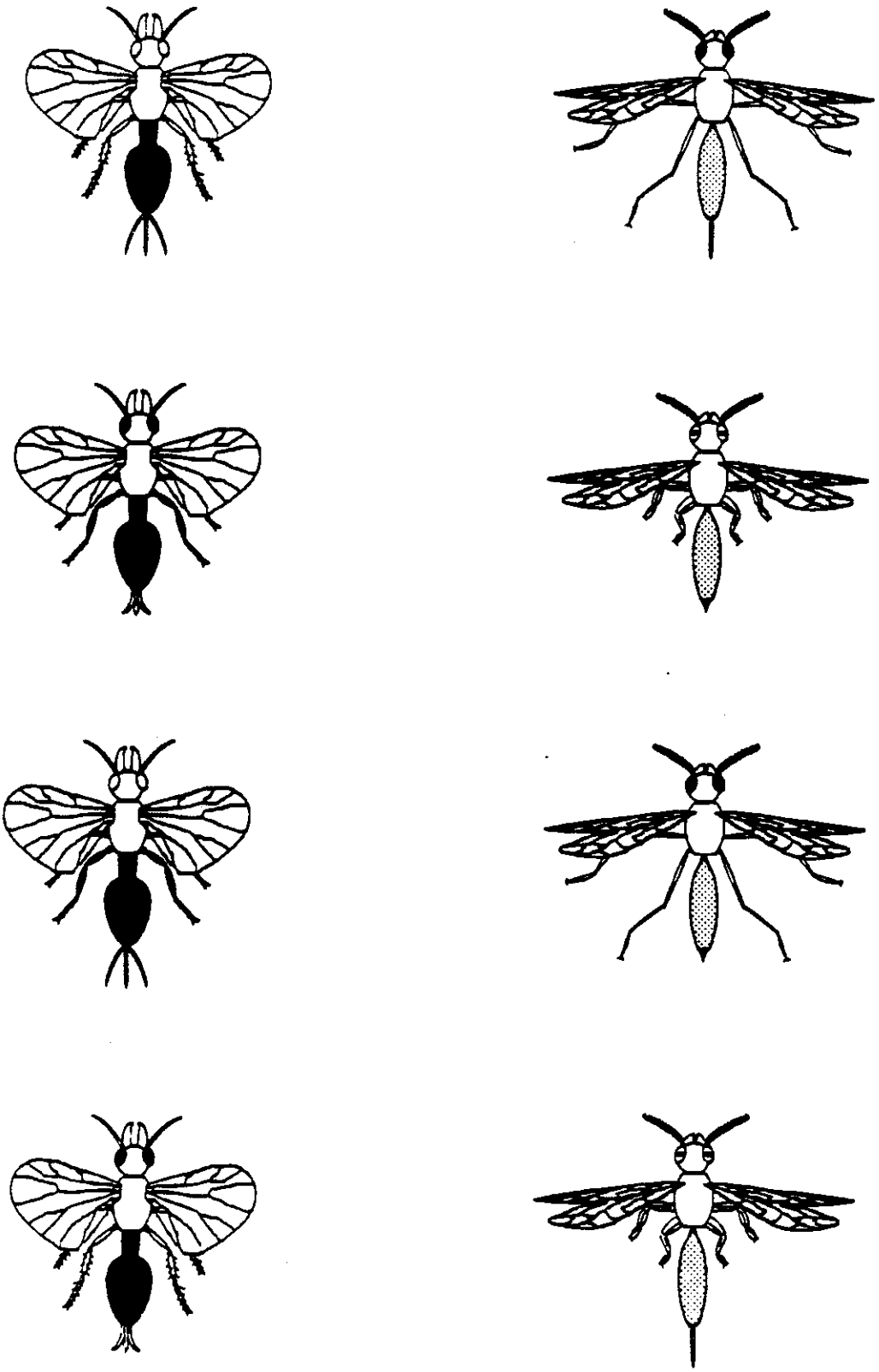
#AFOSR-87-0282

One of the practical tasks people face is how to learn about their environment, in particular, how to categorize and classify the objects and events around them. Practically all experimental research on category learning has studied what is called "supervised learning," wherein a tutor or supervisor teaches a concept to a learner by providing trial by trial feedback regarding the learner's tentative classification of a series of patterns.

We will address a different issue here, namely, how people learn categories when they have no teacher, when left on their own to discover any usable clustering of stimuli that they can. We call the general paradigm "unsupervised learning", because it involves no supervisor or trainer who provides feedback to learners about the current classification. In fact, in our experiments with college students, we never mention categories or category learning. From the subjects' point of view, they are simply trying to memorize each instance or stimulus pattern as it's presented.

We believe that this kind of unsupervised discovery of categories occurs often in real life, perhaps as preverbal children explore their world of perceptual objects, as they learn their language, or whenever pioneers in any unexplored field try to classify the varieties of things that nature serves up to them. Unsupervised category learning occurs in formal school settings often under the name of "discovery learning" wherein students are permitted to explore a given physical domain in hopes that they will stumble upon its underlying structure or principles.

In our research we use as stimuli collections of trees or of insects such as these (Overhead #1). We've composed these pictures on a Macintosh computer;



the insects vary in many physical features, and we define a category of bugs according to which features go together. Thus, these bugs arranged in two columns fall into two categories; they differ in their body shape, color, type of wings, antennae, and front pincers, whereas the eyes, tails, and legs vary within the categories. We can represent the stimuli in abstract binary notation with each instance represented as a row vector, as shown here (Overhead #2). In this illustration, the first 5 of 8 attributes have correlated values of 1 in Category A and 2 in Category B, whereas the last 3 attributes vary randomly within the categories. We refer to the first attributes as predictable or default values of the relevant attributes, and the second as variable values of the unpredictable attributes.

We were interested in two questions. The first question was how to measure category learning in this domain where categories are never mentioned to subjects. The second question was how to arrange sequences of training stimuli shown one at a time in order to speed up category discovery and use.

Turning to the first issue, how might we measure category learning in such situations in which categories are never mentioned? We did this by giving subjects the goal of remembering each of the instances as they were presented one by one. Shown a bug, subjects were asked to list a few of its attributes that would enable them to remember it later, in the sense of being able to pick it out from 3 other similar bugs on a recognition memory test. Subjects were urged to list as few attributes as possible, but ones that would be maximally informative for picking out this bug on a later test. That later test in fact was never given.

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

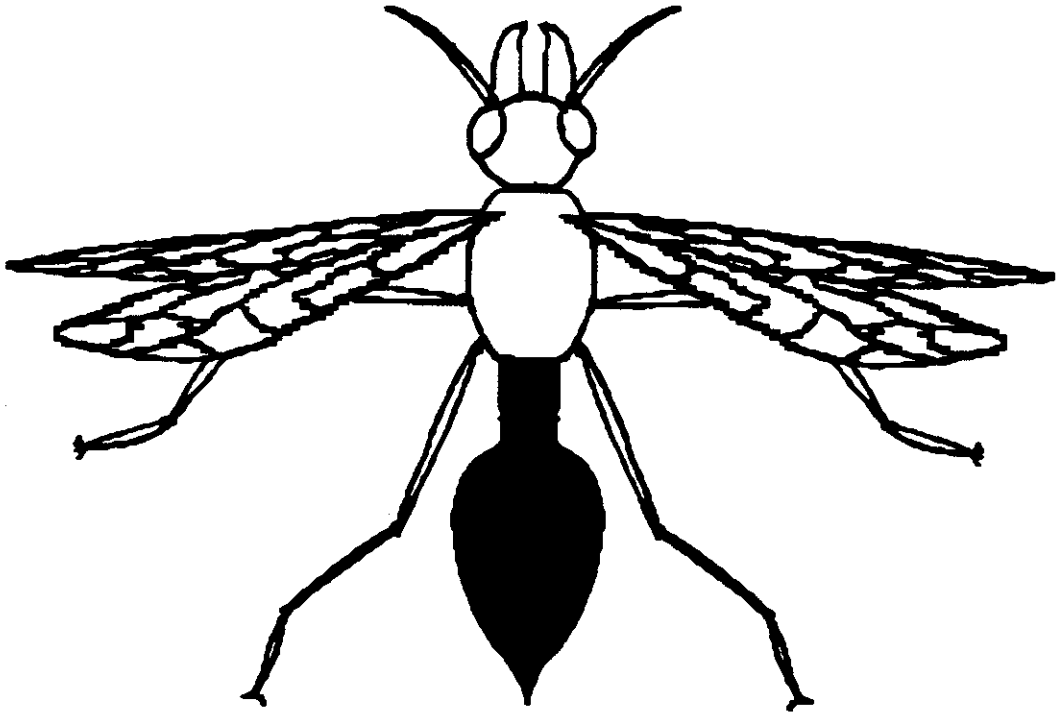
Category "A" : 1 1 1 1 1 x x x
Category "B" : 2 2 2 2 2 x x x

Here is an example (Overhead #3) of what a subject listed for this insect: white eyes, thin wings, black vase-shaped abdomen, 1 stinger, 2 white antennae, and 2 clawlike pincers.

So, what might we expect subjects to do in this task? Subjects might play dumb, and simply list and record in memory for each instance most of its features, leading to a memory representation like that shown at the top of this overhead (Overhead #4). But an intelligent learner ought to follow a "schema-plus-corrections" strategy which leads to a type of memory representation illustrated at the bottom of the overhead. By this strategy, subjects should record a new instance by first noting one or more of its defaults to indicate its category, and then, by listing the unpredictable or variable attributes which would serve to uniquely identify this particular instance within the category. Of course, to follow this strategy, subjects must have first formed a category based on noticing consistently correlated defaults. Looked at from another perspective, however, we can take as an indirect measure of category learning the extent to which subjects stop listing the predictable defaults but increasingly list the unpredictable, variable attributes of the instances.

It turns out that actual learners--at least, college students--come to approximate this ideal pattern. Their performance is illustrated in the next overhead (Overhead #5); these are subjects who first see 16 patterns of one category (call it A) followed by 16 of a second category (called B), followed finally by a mixture of 4 A and 4 B test patterns. The top figure shows the percent of default attributes that subjects list; this drops rapidly from around 60% for the first bug to around 15% by the 10th bug. The listing of default attributes rises abruptly when the first B pattern comes along, since subjects are

#3



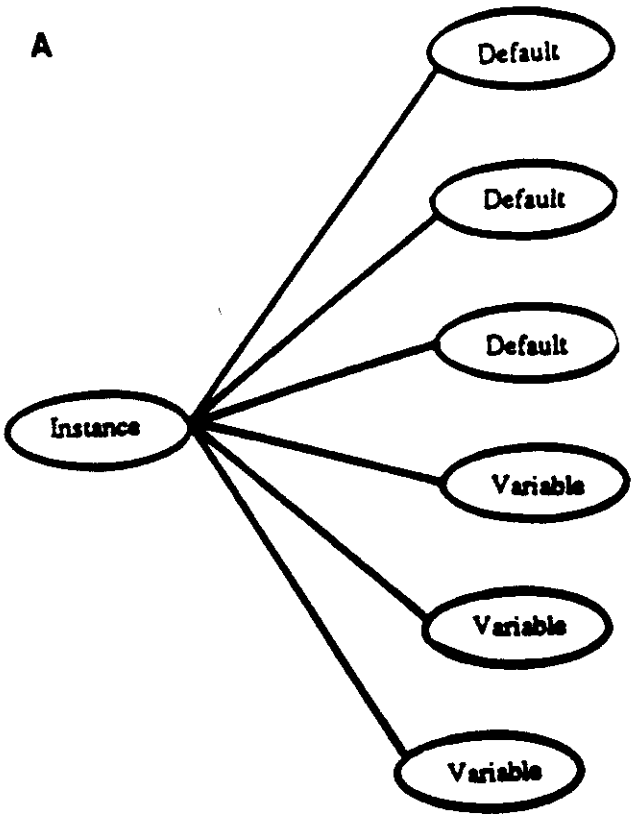
DO NOT WRITE OR MAKE ANY MARKS ABOVE THIS LINE

Write down a short list of this insect's features. List ONLY those features that you'd need to identify this insect on a later multiple-choice test. Imagine that: (1) each feature on your list will cost you **25 cents**, and (2) each misidentification on the multiple-choice test will cost you **1 dollar**.

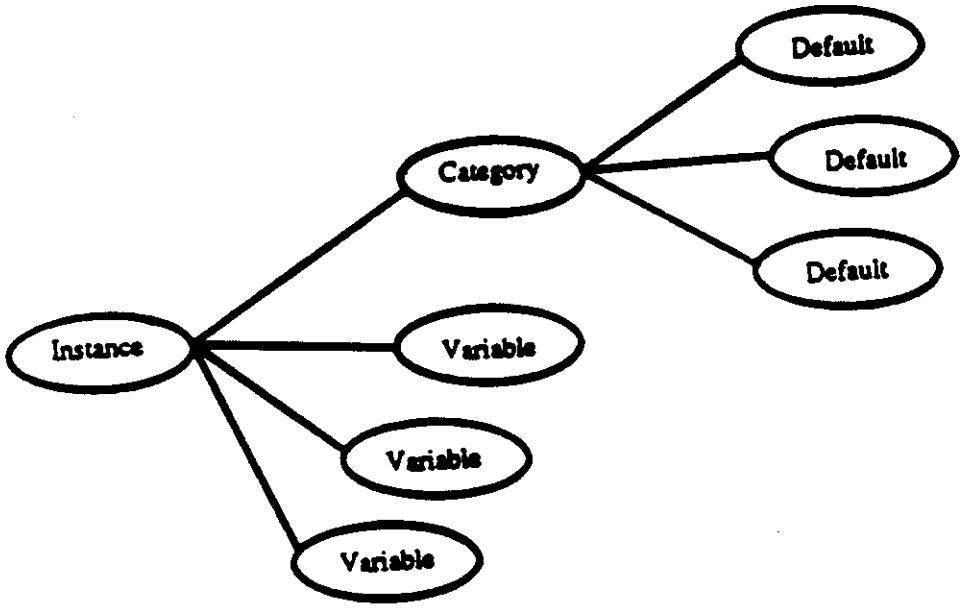
1) White eyes
2) 4 thin wings
3) Black vase shaped abdomen ending in 1 stinger
4) 2 long white antennae + 2 long white clawlike ones

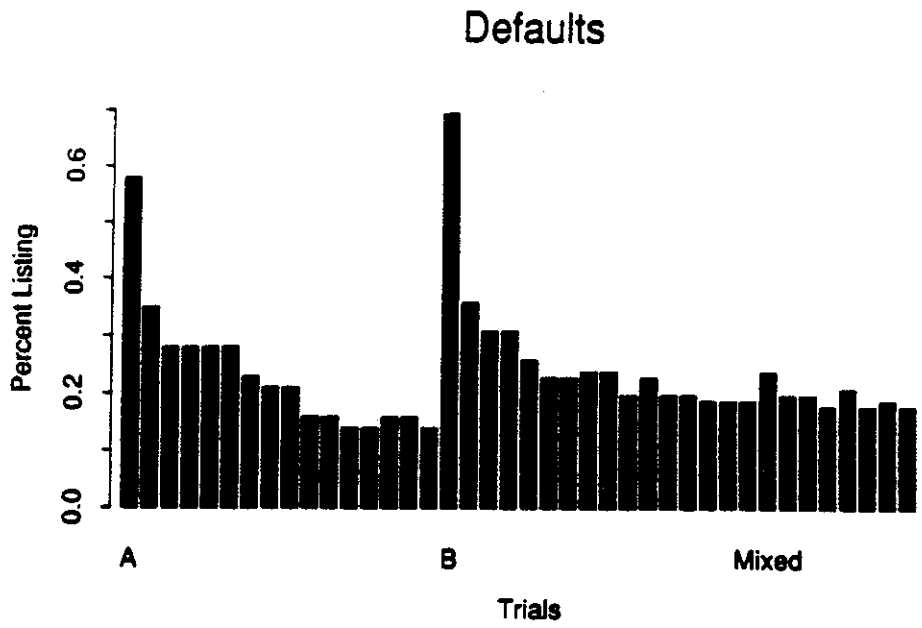
DO NOT LOOK AT ANY OTHER PAGES.

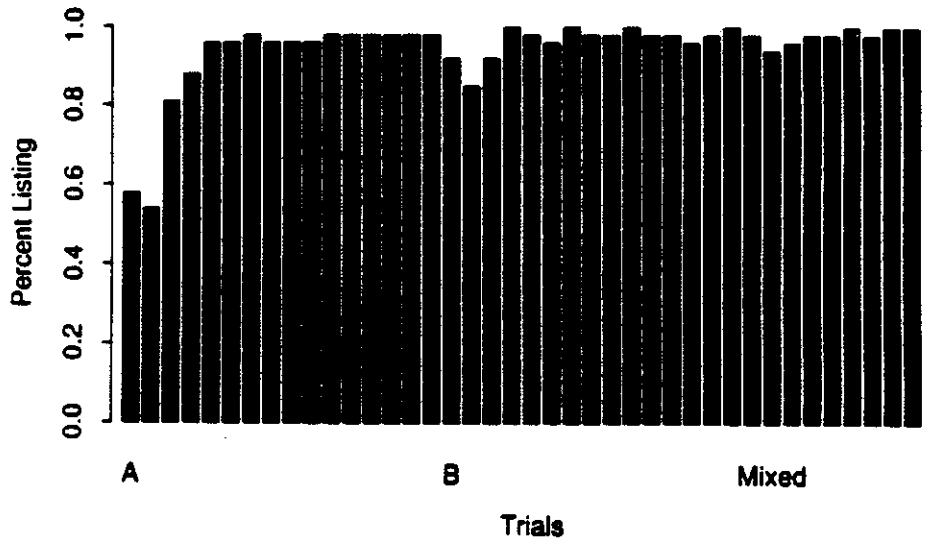
A


Default
Default
Default
Instance
Variable
Variable
Variable

B


Default
Default
Category
Default
Instance
Variable
Variable
Variable

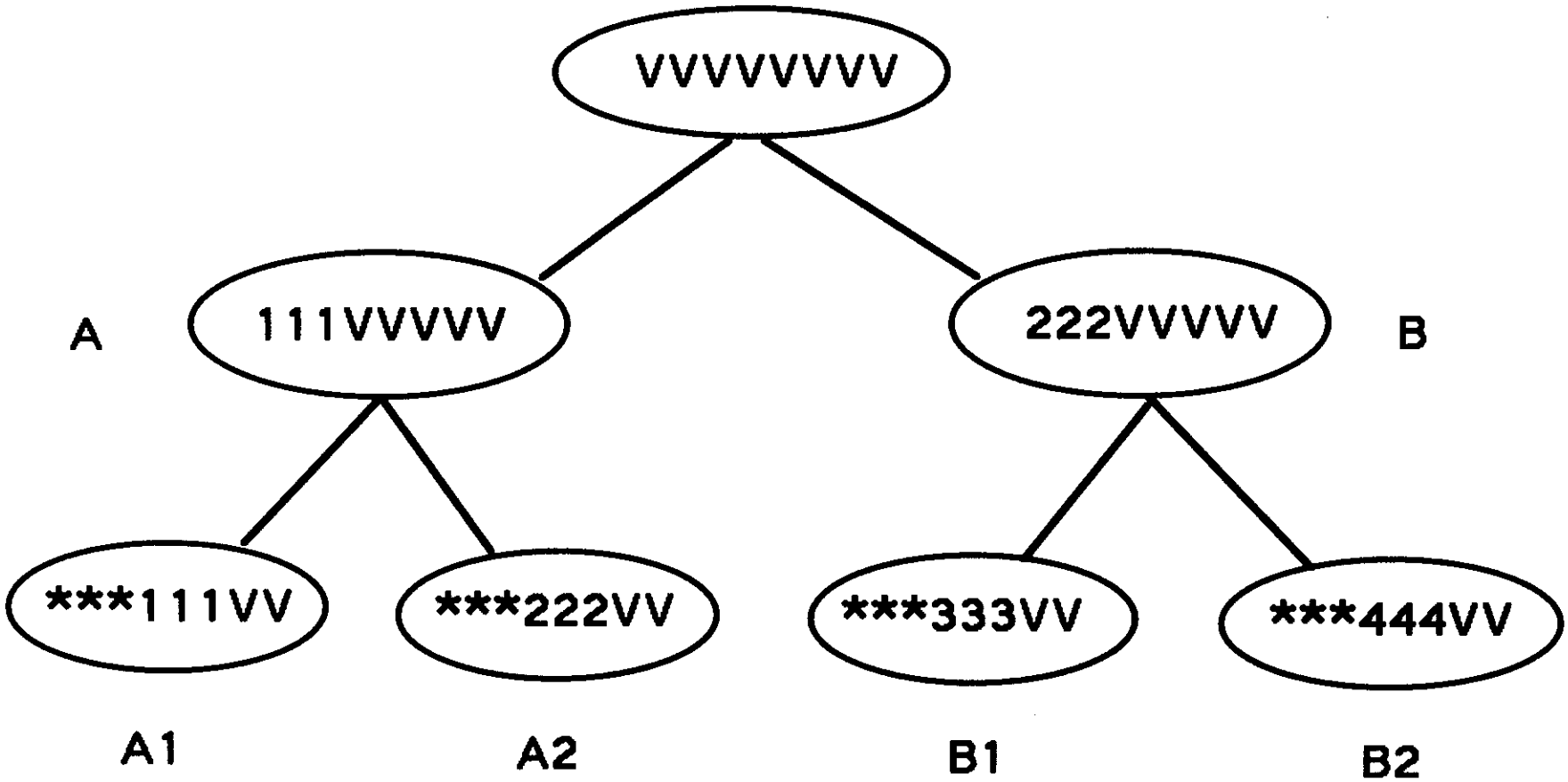

Defaults
Percent Listing
0.0
0.2
0.4
0.6
A
B
Mixed
Trials

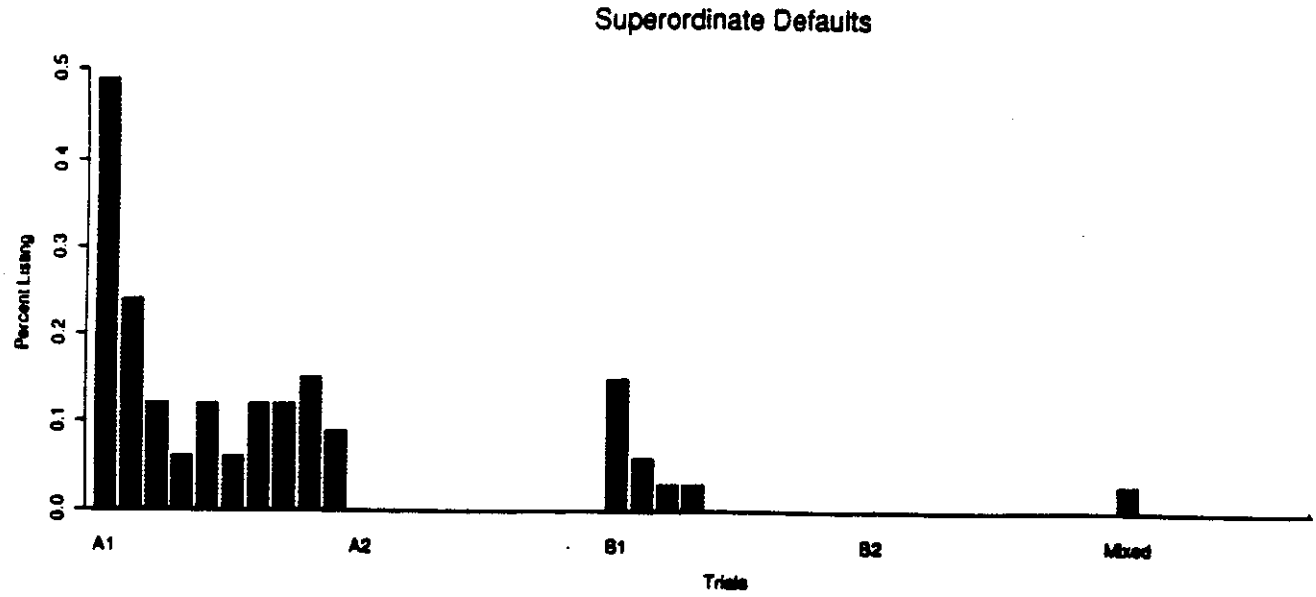

Variables
Percent Listing
0.0
0.2
0.4
0.6
0.8
1.0
A
B
Mixed
Trials

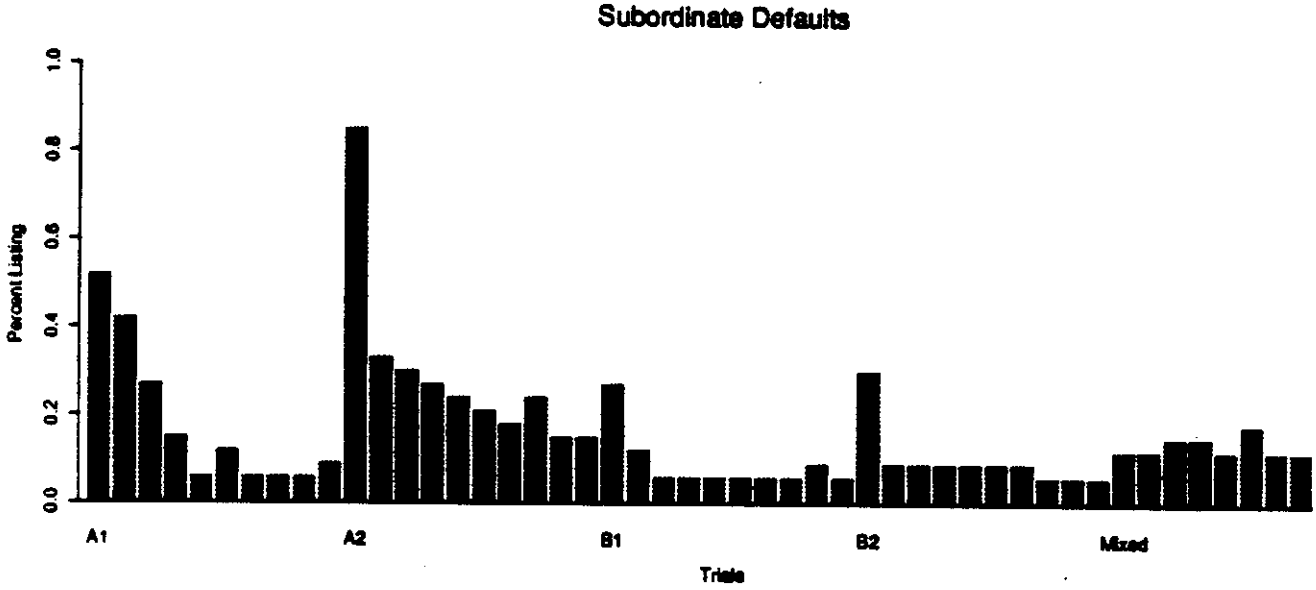
surprised by the novel values of it's default attributes. But here again, listing of default values for the B patterns quickly drops off to near the minimal value as the B-category norms are learned. The minimum number of defaults subjects should record is one out of 5, or 20%, which is exactly where they're performing. It's significant, too, that subjects are not disrupted when, in the final block of trials, they encounter a mixed series of A's and B's. They obviously were able to maintain intact the separate norms for the two categories.

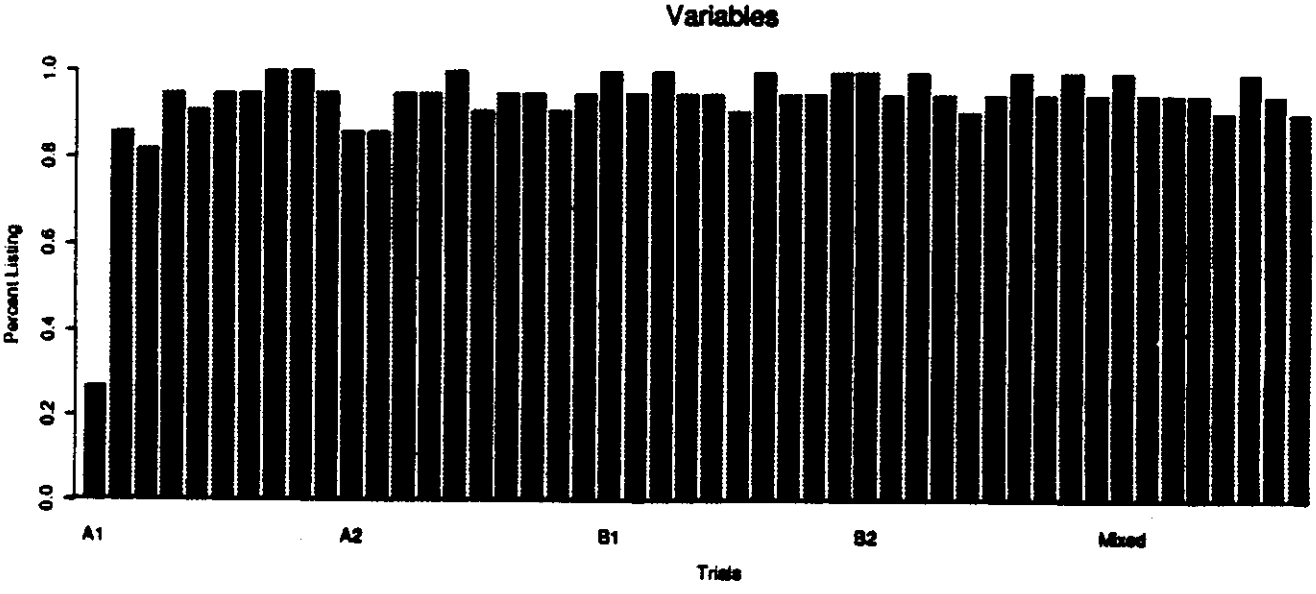
The percent listing of unpredictable, variable attributes at the bottom of the overhead is just the mirror image reflection of the default-listings above. During the inital A-series, subjects learn to identify and list increasingly the variable attributes of the instances; this dips a bit when they hit the first surprising B stimulus when they have to record the new defaults, but listing of variables quickly recovers to a high level throughout the following mixed series of As and Bs. As you can imagine, a useful index of learning is simply the difference in percentage listing of variable minus default attributes.

This same procedure was used to follow subjects learning a hierarchy of nested categories, which is illustrated here (Overhead #6), where V stands for a variable attribute. The first two categories, A1 and A2, share values of 1 on the first 3 attributes but differ on the fourth, fifth, and sixth. The last two categories, B1 and B2, share values of 2 on the first 3 attributes but likewise differ on the fourth, fifth, and sixth attributes. These were shown to subjects, in blocks of 10 instances with the blocks in the order A1, A2, B1, B2, all ending with a mixed test block of 2 instances from all 4 subcategories. Here are the attribute listing data (Overhead #7). The top graph shows the rapid decline in listing of the superordinate (the first three) defaults, but it pops up again at the first B


VVVVVVVV
A
111VVVVV
222VVVVV
B
***111VV
***222VV
***333VV
***444VV
A1
A2
B1
B2


Superordinate Defaults
Percent Listing
0.0
0.1
0.2
0.3
0.4
0.5
A1
A2
B1
B2
Mixed
Trials


Subordinate Defaults
Percent Listing
0.0
0.2
0.4
0.6
0.8
1.0
A1
A2
B1
B2
Mixed
Trials

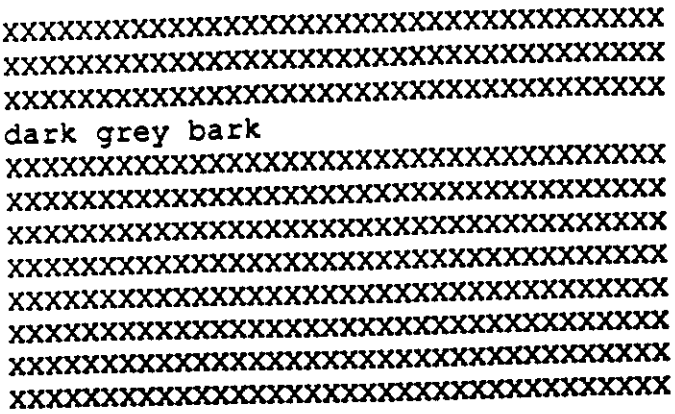

Variables
Percent Listing
0.0
0.2
0.4
0.6
0.8
1.0
A1
A2
B1
B2
Mixed
Trials

stimulus when these values are changed. The middle graph is for the subordinate defaults-- attributes 4, 5, and 6; these decline over the first series of A1 patterns, then pop up when the first instance of the novel A2 subcategory is seen, and pop up again as each new subcategory makes its appearance. However, all the superordinate defaults are carried over without fail as new subordinate categories are created to decribe the new instances. By comparison, the fully variable features, shown in the bottom graph, rise up quickly and continue to be listed around 95% throughout. So these data show excellent learning of a hierarchy of categories of insects defined by common features.

Another indirect index of such learning we have explored involves limiting the amount of time subjects can inspect a verbal description of a species of trees while trying to memorize it (Overhead #8). They are allowed to examine the attributes one at a time on the computer screen. They can move around among the attributes, and we record how much of the time they invest in studying default versus variable attributes. After studying each instance, their memory for it is immediately tested. As expected, subjects quickly learn the defaults, so they spend progressively less time inspecting them, but nonetheless recall them very well; on the other hand, they spend a progressively higher percentage of their time studying the variable attributes, so that their memory for those features also improves. I'll not have time here to show you any of that data on self-selected study times, although it is very orderly and regular.

I will turn now to describing a procedural variable we've studied which has a major impact on how much subjects learn from exposure to a set of A and B patterns. This variable is simply the sequence or order in which the two sets of patterns are presented to subjects. We were not prepared for the huge effect this

a. Aralia



Press INS or DEL to see next item

---

b. Aralia

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

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

---

c. Aralia

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

INCORRECT

Arrow indicates correct choice

Press RETURN to go on

variable had on our data, namely, that category learning is far, far easier if subjects see a long block of instances all of one type before they ever see instances of a second type. Compared to a randomly intermixed series, the advantage in learning produced by blocking is just enormous.

You can get an intuitive feel for the difference in difficulty here by perusing the first 10 instances (in the rows) which mix 5 A-patterns with 5 B-patterns in random order in this overhead (Overhead #9); your job is to find out what values of which attributes are correlated. It's very hard to do; moreover, our subjects weren't allowed to examine all 10 patterns laid out before them at one time as you can see here, they could only see one instance at a time. Iin addition, our subjects did not have the goal of looking for correlated features underlying categories.

The structure of the patterns becomes more obvious and easily learnable if the same 10 instances are presented in a block of 5A's, then 5B's, as illustrated at the bottom. Here, looking down the columns, you can see that the A's have a value of 1 in attributes 2, 4, 7, and the B's have a value of 2 in those attributes.

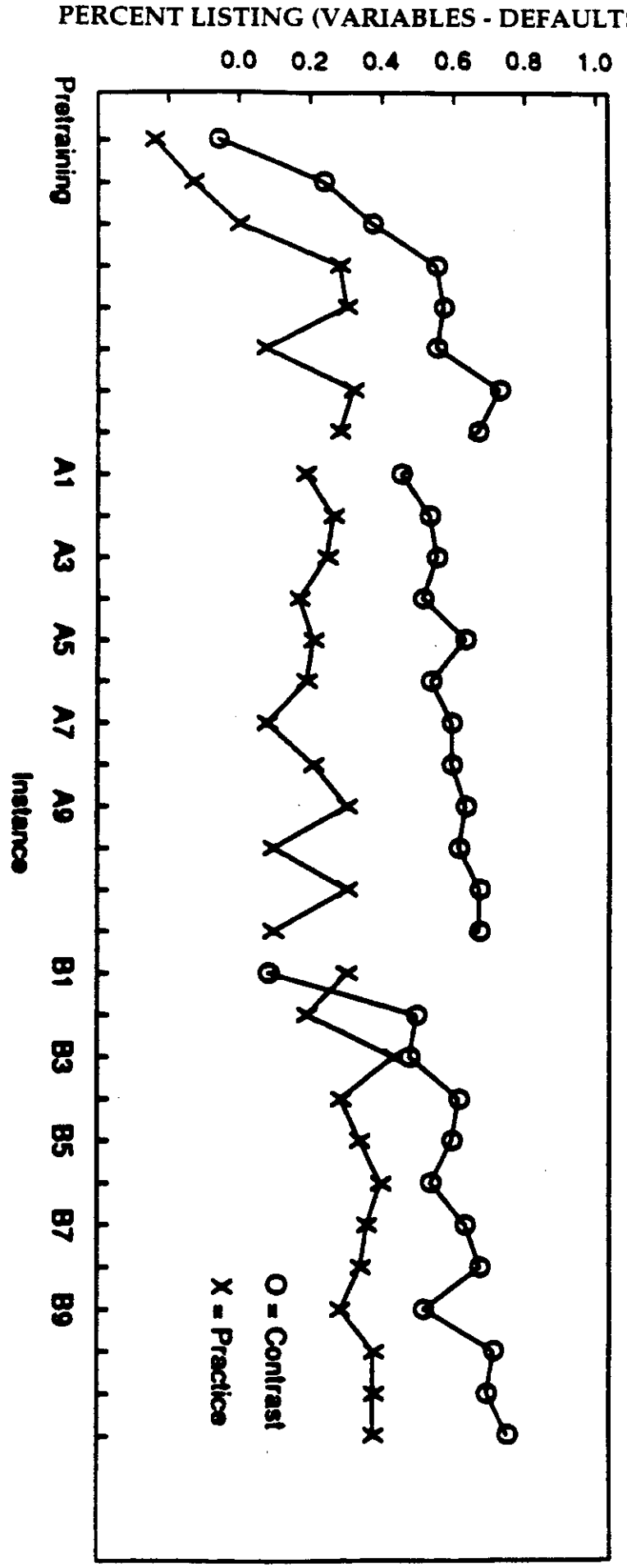
It's easy to demonstrate this principle, that category learning is facilitated by blocking instances. One demonstration is shown in this overhead (Overhead #10) comparing learning of two groups after they were pretrained with 8 instances. In pretraining, the Practice group (the Xs) saw 4 A's mixed in with 4 Bs, whereas the Contrast group (the O's) saw a block of 8 A's. These two groups then received a mixed series of 12 A's and 12 B's, but we've plotted the A and B trials separated in this graph. The learning measure is the subjects' preference for listing variable over default attributes.

	ATTRIBUTES							
INSTANCES	1	2	3	4	5	6	7	8
1	2	1	2	1	1	2	1	1
2	1	2	1	2	2	2	2	1
3	2	2	2	1	2	2	2	1
4	1	1	2	2	1	1	1	1
5	1	2	1	2	2	1	2	2
6	1	1	2	2	1	2	1	1
7	2	1	1	1	1	1	1	2
8	2	2	1	1	2	1	2	2
9	1	1	1	2	1	2	1	2
10	1	2	2	1	2	1	2	1

MIXED SEQUENCE

	ATTRIBUTES							
INSTANCES	1	2	3	4	5	6	7	8
1	2	1	2	1	1	2	1	1
2	1	1	2	2	1	1	1	1
3	1	1	2	2	1	2	1	1
4	2	1	1	1	1	1	1	2
5	1	1	1	2	1	2	1	2
----------	----	-----	-----	-----	-----	-----	-----	-----
6	1	2	1	2	2	2	2	1
7	2	2	2	1	2	2	2	1
8	1	2	1	2	2	1	2	2
9	2	2	1	1	2	1	2	2
10	1	2	2	1	2	1	2	1

BLOCKED SEQUENCE


PERCENT LISTING (VARIABLES - DEFAULTS)
0.0
0.2
0.4
0.6
0.8
1.0
Pretraining
A1
A3
A5
A7
A9
B1
B3
B5
B7
B9
Instance
O = Contrast
X = Practice

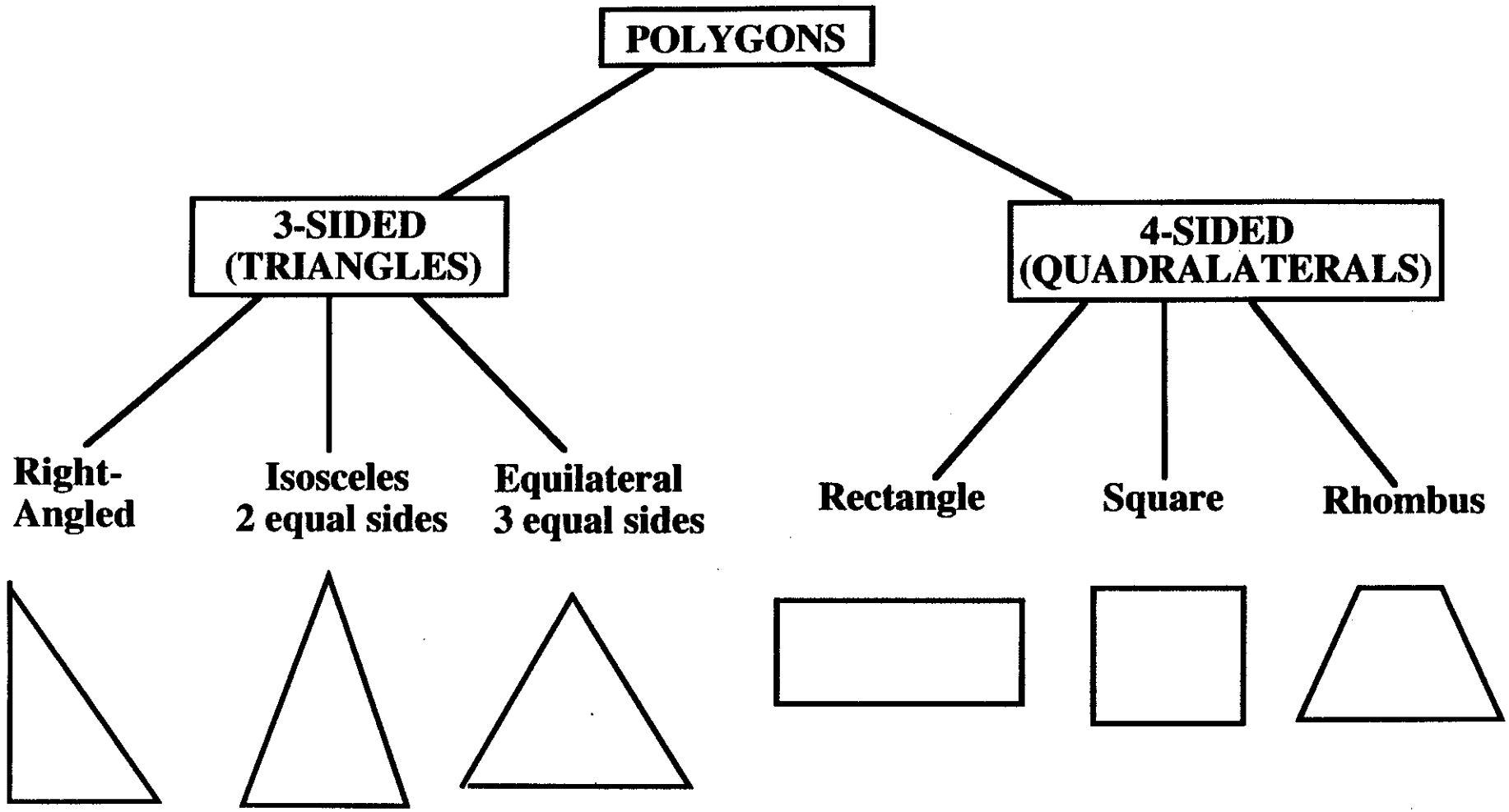
It's obvious that the Contrast subjects, who start out seeing enough A's in a row to acquire very confident norms, continue to do well even after they are surprised by the first B pattern; they consistently outperform the Practice subjects, whose pretraining with a mixture of A's and B's seems to have locked them into a very nonoptimal performance, barely above that of a random uncorrelated control condition. Paradoxically, the Contrast condition shows that by eliminating the B trials in Pretaining, we enhance the learning of B stimuli in the later series.--- sort of a reverse practice effect. The Practice subjects apparently see so much variability in their pretraining mixture of A's and B's that they give up any attempt to see structure in the collection of patterns. [Overhead off]

This difference in learning between blocked vs. mixed series is very interesting theoretically because the result contradicts most of the standard theories of unsupervised learning. For example, many clustering algorithms expect best performance when the model sees many contrasting examples in alternation rather than a block of one type. And connectionist models that use autoassociation to learn interfeature correlations predict either no difference due to sequencing stimuli or predict catastrophic retroactive interference with the blocked sequence.

The results instead point to a discrete process by which subjects invent a category, then acquire norms or defaults within that category that are sufficiently stable that subjects are able to be surprised by an unexpectedly large departure from those norms when they encounter the first instance of the alternate category. That surprise leads them to set up a new category and to

begin learning its norms in a manner segregated from the norms learned about the first category. We've developed a simulation model that does just that, but I won't bother presenting it in this setting.

So those are the results I wished to present. The additional question for this conference is whether there are any practical applications of the result. I suppose some applications could certainly arise in educational settings where students are learning to distinguish between members of different categories, such as different animals, plants, flowers, airplanes, ships, or styles of residential architecture. In these cases, we have some chance of describing objects in terms of lists of features. Another example might be for students learning to identify geometric figures, such as regular polygons, as shown here (Overhead #11). Polygons can be divided according to their number of sides, with names provided for some that have special features. Thus, triangles can be classified as right triangles, isosceles, or equilateral depending on special features. If we wanted students to discover these classes for themselves, we could show them example triangles arranged in an order that was either blocked or random across the subcategories. Presumably they would discover the classes more quickly if they observed the examples in a blocked fashion. I think the blocking strategy would work well with acquiring expertise in wine tasting: I can imagine that prospective wine tasters would learn to discriminate the wines more quickly if they tasted a collection of chardonnays, then rieslings, then sauvignon blancs in blocked fashion rather than tasting them in a mixed up sequence. You will all have the opportunity to test out this prediction at the receptions during this conference.


POLYGONS
3-SIDED (TRIANGLES)
4-SIDED (QUADRALATERALS)
Right-Angled
Isosceles 2 equal sides
Equilateral 3 equal sides
Rectangle
Square
Rhombus

Another application of the blocking idea would be general advice to unsupervised learners for when they explore an uncharted domain, especially if they have some control over the order in which they see examples. The advice is to avoid covering too much territory too quickly lest you get overwhelmed by the variability. Rather, it is better to start out slowly by exploring only relatively small variations of aspects of a given type of object--say, types of leaves on plants. In this way, you can learn a confident set of norms for that one type. Thereafter, using that as a firm foothold for classifying the domain, you can seek out a large contrasting type to learn next, and begin exploring small variations around that contrasting class. Our results suggest that, when it can be implemented, this strategy should produce fairly rapid discovery learning. That is, at least, one of the practical lessons I draw from this otherwise theoretical result.

Thank you for your attention.