CHAPTER
19

Category Learning as Schema Induction

John P. Clapper
California State University–San Bernardino

Categories are essential for practically all aspects of human cognition, and an enormous amount of research has been devoted to understanding how they are learned and represented in memory (see, e.g., Murphy, 2002; Smith & Medin, 1981). In this chapter, I describe a program of research on category learning that Gordon Bower and I began a number of years ago, when I was a graduate student in his laboratory at Stanford University, and which I have continued to extend and develop over the succeeding years. This research began by investigating how people use category knowledge (schemas) to guide attention and organize memory, and later focused on the basic mechanisms by which such categories are discovered and learned. Although category learning has been a traditional focus of research within cognitive psychology, we have taken a rather non-traditional approach in our own investigations of this area. My primary goal in this chapter (besides paying tribute to Gordon) will be to convince readers of the value of this non-traditional approach.

It is important to note that most category learning research has been carried out within a standard conceptual framework or *paradigm* that describes what categories are, what they are for, and the kinds of situations in which they are normally acquired. Within this paradigm, category learning typically is regarded as a form of

discrimination learning and is investigated using standard discrimination-learning procedures (e.g., Kling & Riggs, 1971). In a discrimination-learning task, the participant is presented with a series of stimuli and is taught (using corrective feedback) to respond differently to different classes of stimuli (e.g., Bruner, Goodnow, & Austin, 1956; Medin & Schaffer, 1978; Posner & Keele, 1968, 1970). Within the category-learning literature, this task is usually called classification learning or *supervised* learning (Michalski & Stepp, 1983); the latter term refers to the fact that discrimination feedback is provided by an external "supervisor."

Although the discrimination-learning paradigm has proved itself useful over many years of research, I will argue that it has also led to a somewhat limited and misleading view of categories and category learning (see Markman & Ross, 2003 for related arguments). Perhaps the most fundamental limitations of this paradigm are due to the way that categories are defined in discrimination-learning tasks. In this type of task, a category is simply an equivalence class defined by a common label or response, and "knowing" a category means nothing more than being able to assign instances to that category—in other words, being able to discriminate examples of that category from examples of the other category or categories currently being shown. Because members of a category need have nothing more in common than a shared label, categories can be defined arbitrarily, that is, any arbitrary collection of stimuli may be defined as a valid category simply by assigning them all the same label. And because categories are defined only in terms of discriminative responding, one need learn only the minimum features required for correct discrimination to be regarded as knowing a category (Markman & Ross, 2003), even if that means *not* learning most of the consistent structure within that category (e.g., even diagnostic features need not be learned if they are redundant with other diagnostic features).

This is a rather odd definition relative to our everyday way of thinking about categories, in which knowing a category means knowing what members of that category are *like,* not just the ability to distinguish category members from members of a specific contrast category (Chin-Parker & Ross, 2004). As we see later, this definition leads to a tendency to ignore a number of important issues relating to how categories are learned, represented, and used. In addition to defining category learning in this narrow way, the discrimination-learning task also includes such artificial features as explicit instructions to classify each stimulus, a convenient set of predefined category labels to choose among, and corrective feedback on every trial. One might reasonably ask whether these features are really necessary for people to learn categories, and if not, whether a simpler task might not provide a more appropriate model learning situation.

In principle, the *minimal* category-learning task would be one in which the participant simply receives a series of training instances that are *potentially* divisible into separate categories, with no instructions to search for categories and no predefined category labels or trial-by-trial feedback. As they examine each stimulus, the person's current knowledge state would presumably be altered in some way by that experience. If it were possible to track those changes unobtrusively over time, it might be possible to find out whether they discovered the categories on their own and to trace the course of this learning over trials. Any learning in

this kind of task would obviously be *unsupervised,* because no category-level feedback is provided, as well as *incidental,* because the person is not asked to search for categories and would therefore have to notice and learn any categories on their own initiative.

In this chapter, I refer to this as a "schema induction" task; I also refer to the particular approach that goes along with it as the schema induction framework or paradigm. As I explain in greater detail later, this task differs from a discrimination-learning task in several ways; most important is the fact that learning is defined in terms of knowing as many features as possible within a given category, not merely being able to tell members of different categories apart. One benefit of this re-framing is that it facilitates the investigation of interesting questions that either do not arise or are difficult to study within the standard discrimination-learning approach.

EARLY RESEARCH ON SCHEMA APPLICATION

The original impetus for this approach came not from traditional studies of supervised categorization, but rather from studies of memory for text and other forms of discourse, and the role that organized knowledge or *schemas* play in this kind of memory (e.g., Bartlett, 1932; Minksy, 1975; Rumelhart & Ortony, 1977; Schank & Abelson, 1977). Here, the term *schema* is used to refer to an internal model or representation that contains knowledge about a specific category. In theory, research on schema-based memory is directly relevant to understanding categories because it investigates how general, category-level knowledge is used to remember specific instances or situations. However, such issues are outside the purview of standard discrimination-learning procedures, in which people are only asked to classify the stimuli rather than using category knowledge to predict or reconstruct their features.

Part of our interest in these issues stemmed from Gordon's prior research on the role of schemas in memory for text (e.g., Bower, Black, & Turner, 1979; Belleza & Bower, 1981). Historically, most proposals about schemas and memory have been variations of the "schema-plus-corrections" (S + C) theory put forward several decades ago (e.g., Attneave, 1954; Oldfield, 1954). This theory begins by noting that many of an object's features are predictable from its schema, which specifies the structure common to the category as a whole. Therefore, the most economical way to represent an object in memory is to include only features ~~those~~ that are not inferable from the general schema, while referring to the schema itself for those that are. Economy of storage thus ~~is~~ gained by eliminating unnecessary redundancy from the memory trace. This framework featured prominently in early research on memory for meaningful text. One of the main findings of this research is that people are likely to falsely recall or recognize events from a script-based story that were in the underlying script (event schema; see Schank & Abelson, 1977) but were not actually stated in the text (e.g., Bower, Black, & Turner, 1979). Graesser and coworkers (e.g., Graessar, 1981; Graesser, Woll, Kowalski, & Smith, 1980) showed that people's ability to discriminate whether or

not an event had been explicitly stated in a text declined as a function of the typicality of that event within the script or schema, such that highly typical events showed nearly zero memory discrimination following a 30-minute retention interval. To explain these findings, Graesser proposed a "schema-pointer-plus-tags" (SP + T) model of memory. The SP + T model proposed that schema-based texts were encoded by (a) creating a "pointer" to the general script or schema referred to in the text and (b) encoding specific traces or "tags" for any events that were not highly typical or expected within that schema. Bower, Black, and Turner proposed a similar model that assumed that readers create a partial memory trace of the story and rely on their script-based knowledge to fill in missing details. Both models were closely related to the original S + C framework in relying on a general schema to reconstruct the typical or expected features of a given instance.

Whereas these theories emphasized the storage and retrieval aspects of schemas, another theory, the attention-elaboration hypothesis (e.g., Belleza & Bower, 1981; Bobrow & Norman, 1975) argued that at least some of the advantage in memory discrimination for atypical or deviant events may have been due to differences in the amount of attention they received at the time of encoding. For example, Belleza and Bower showed that people spend more time reading atypical compared to typical statements during a prose comprehension task.

Gordon and I later proposed a theory that combined the S + C theory and the attention-elaboration hypothesis within a single framework (Clapper & Bower, 1991; Fig. 19–1). The theory assumes that people attempt to categorize each stim-

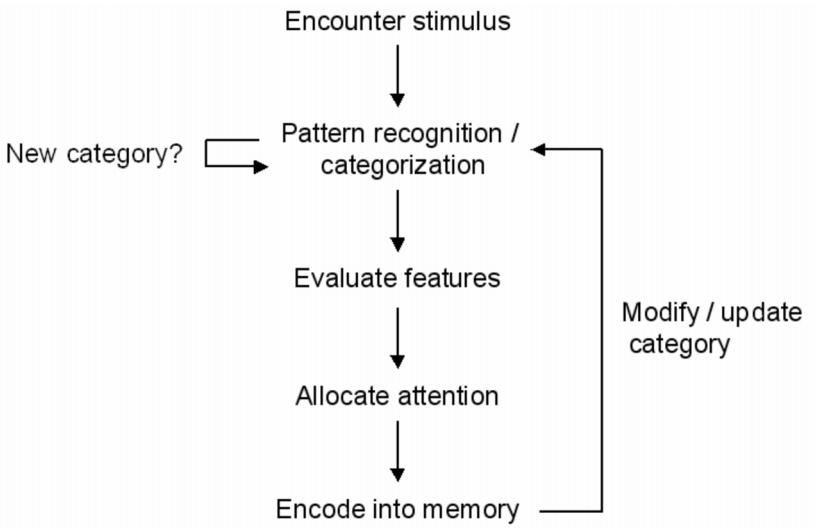

Figure 19–1. Instance- and category-learning model based on Clapper and Bower (1991).

ulus they encounter as part of their normal pattern recognition process. Features that are expected or typical within that category are considered uninformative because they are redundant with the general schema; therefore they should receive little attention during encoding. Unpredictable or surprising features are considered highly informative and should receive a much greater share of the learner's attention. The result of these attentional/encoding biases is that the distinctive features of an instance should tend to be associated directly with that instance in memory, whereas features typical of the general category should tend to be associated mainly with the schema and only weakly, if at all, with the individual instance (Fig. 19–2).

This kind of memory organization has important implications for memory retrieval. A large number of experiments on the so-called "fan effect" (e.g., J. R. Anderson, 1976, 1983; J. R. Anderson & Bower, 1974) have shown that the more items that are associated with a given concept in memory, the slower people are to retrieve any of those items. This suggests that the more features that a person has directly associated with an instance in memory, the longer it should take them to verify any of these features for a speeded recognition task. However, if category-typical features are not associated directly with the instance, but rather with the category schema (as in Fig. 19–2), then the situation becomes more complex. In this case, the number of category-typical features should have no effect on the time to verify distinctive features of an instance, because the category features are not directly associated with that instance in memory. The opposite, however,

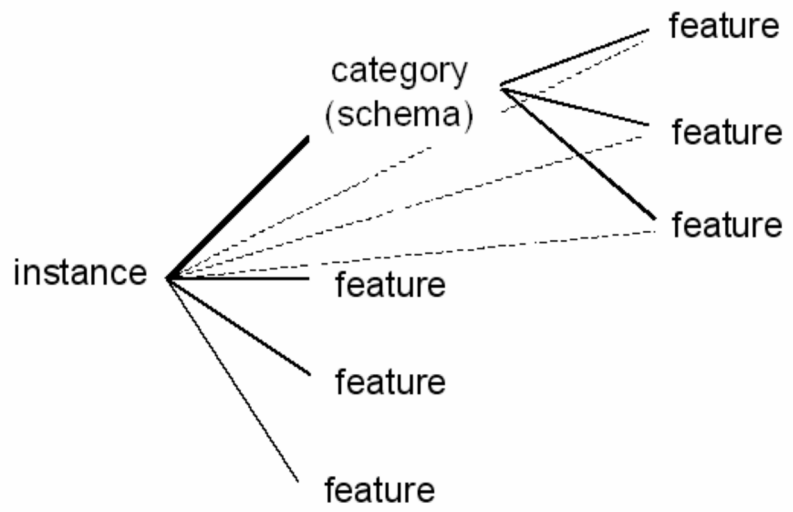

Figure 19–2. Schema-plus-corrections memory organization resulting from the attentional assumptions of the Clapper and Bower (1991) model.

would not be true; if we think of category membership as a feature of an instance, then the more distinctive features associated that that instance the longer it should take to retrieve its category membership and the typical features associated with that category.

An experiment that we conducted to test these ideas consisted of three phases (see Clapper & Bower, 1991, for a more complete description). In the first phase, people learned several different categories, each presented in the form of a verbal feature list (the categories were types of astronomical stars and the features were their chemical constituents). In the second phase, they learned several instances of each category. Each instance contained all the elements possessed by its parent category, as well as one or more additional elements distinctive to that particular instance. The number of features (fan) associated with each category of stars (2, 3, or 4) and the number of distinctive features associated with each individual star (1 or 3) were varied orthogonally. In the third phase, participants were given a speeded recognition task in which they had to verify whether or not particular instances possessed particular features or belonged to a specific category.

The results were strongly consistent with our hypothesis that people were storing category information separately from information about specific instances. The more distinctive features possessed by a given instance, the longer it took people to verify either distinctive or category features of that instance. By contrast, the number of category features had no effect on verification times, suggesting that these features were stored separately in memory, in accordance with our modified S + C model.

LATER RESEARCH ON SCHEMA ACQUISITION

In the experiment just described, people were directly taught several categories prior to encountering specific instances of those categories. As a next step, we wanted to extend our findings to situations in which people learned categories inductively, via exposure to individual instances. In addition, we wanted more direct evidence for our assumption that the S + C memory organization detected in the last experiment was actually a side effect of events that occurred at encoding.

Schema induction tasks. We developed two new tasks that proved especially useful in pursuing these goals. Both are good examples of the type of schema induction task described at the beginning of this chapter. The first (Clapper & Bower 1991, 1994) was an "attribute-listing" task that involved presenting a series of instances from one or more categories and asking participants to list some of the features of those instances. (The stimuli were pictures of fictitious insects that varied along several dimensions; see Fig. 19–3). They were asked to include only those features that they believed would be useful for identifying that specific instance while omitting redundant features that provided no identifying information. The stimulus sets were designed so that members of a given category shared

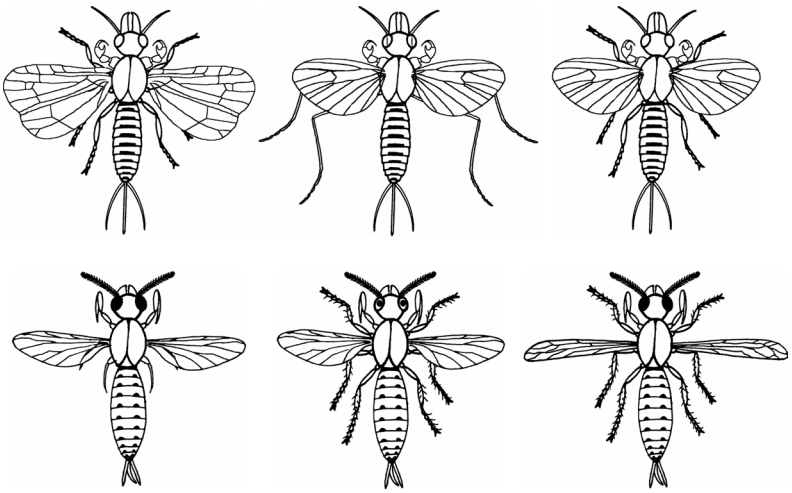

Figure 19–3. Insect stimuli similar to those used in the attribute-listing experiments of Clapper and Bower (1991, 1994). Instances of one category are shown on the top row and instances of another category are shown on the bottom row.

a large proportion of their features, while also varying along several dimensions (Fig. 19–4). Learning was indicated by a decrease in the listing of consistent values, as these were redundant with the instances' category memberships and thus were uninformative for picking out specific instances. At the same time, listing of variable dimensions increased because these were informative for distinguishing among different instances within a category. Listings of variable dimensions minus those of consistent dimensions could be plotted to reveal "learning curves" for the presented categories (Fig. 19–5).

Our second schema induction task employed an immediate-memory procedure. In this task, participants were again presented with a series of training instances from one or more different categories. This time, the stimuli were verbal feature lists describing, for example, fictitious trees, insects, or people (Fig. 19–6). The features in each list were masked by a string of *X*s (Fig. 19–6a), and the participant could uncover and view only one feature at a time (by pressing a designated "up" or "down" key). The computer recorded the amount of time the person spent studying each feature in the list. The list remained on the screen for a preset period of time (say, 24 sec), after which it disappeared and was followed by a series of forced-choice recognition tests (Figs. 19–6b and 19–6c; later versions of this task have also employed cued-recall tests). Following these tests, another instance list would appear and the process would repeat, until all the training stimuli had been shown.

There are two measures of learning in this task. As people learn the categories, they should learn to spend less time attending to the consistent features of each

		Dimensions								
	Example	Shape	Markings	Jaws	Forelimbs	Tails	Antennae	Legs	Wings	Eyes
Category A	A1	1	1	1	1	1	1	1	1	2
	A2	1	1	1	1	1	1	1	2	1
	A3	1	1	1	1	1	1	2	1	2
	A4	1	1	1	1	1	1	2	2	1
	etc.	1	1	1	1	1	1	X	X	X
Category B	B1	2	2	2	2	2	2	3	3	3
	B2	2	2	2	2	2	2	3	3	4
	B3	2	2	2	2	2	2	4	4	3
	B4	2	2	2	2	2	2	4	4	4
	etc.	2	2	2	2	2	2	Y	Y	Y

Figure 19–4. The design of the stimulus sets in the attribute listing experiments of Clapper and Bower (1991, 1994). Notice that each category has six dimensions with consistent values and three dimensions with variable values.

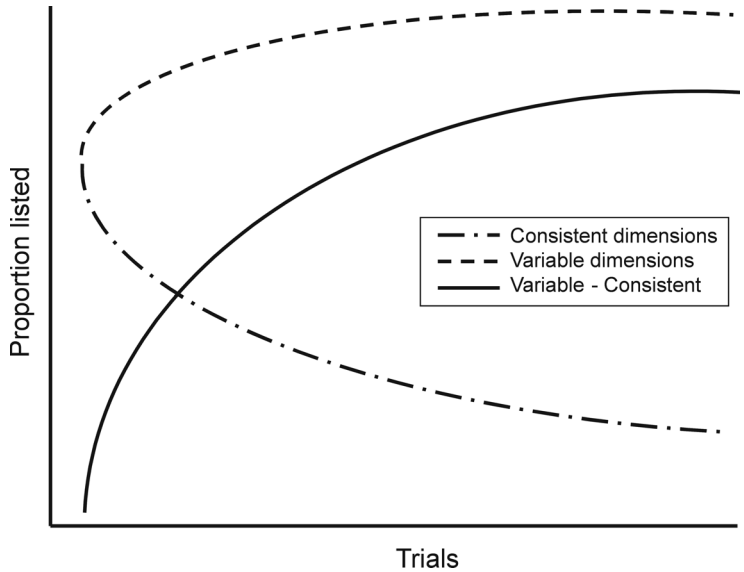

Figure 19–5. Indicators of category learning in the attribute-listing task. The proportion of consistent dimensions listed should decline with trials, whereas the proportion of variable dimensions should increase over trials. When the former is subtracted from the latter, the resulting difference score provides an overall index of feature learning.

Figure 19–6. Exemplar-memory test procedure (structure of a single trial). Each trial involves presenting a study list and several recognition tests. The experiments discussed in this chapter used 30 to 60 trials. (a) Training instance presented on a single trial of the study–test procedure. (b) Example of a multiple choice recognition test for the training instance shown to the left (several such tests were given per trial). (c) The participant has entered a response to the recognition test, and the correct answer is shown. (Adapted from Clapper ~~7~~ & Bower, 2002).

instance (because these are redundant with category membership) and more time attending to their variable features. Subtracting the mean study time for consistent features from that of variable features yields a difference score that can be plotted to reveal learning curves for each category, similar to the difference score used in the attribute listing measure (Fig. 19–7). The memory tests following each instance provide a second measure. As people learn the consistent features of the categories, their memory for these features should improve (Fig. 19–7). Thus, memory performance can also be used to trace learning curves for the categories over trials.

These tasks differ from a discrimination learning task in several ways. Most important, they define (measure) category learning *indirectly,* as a function of how well the person learns the individual features within the category, rather than directly by having them sort or classify the stimuli. This indirect measurement permits *incidental* learning: Participants need not be informed about the categories they are expected to learn, because their memory, study times, or attribute listings will automatically indicate whether they notice such categories on their own. Learning is also assessed *continuously* in these tasks; that is, the output measures provide what amounts to a trial-by-trial record of changes in the learner's knowledge state.

One of the main advantages of these tasks is that they allow us to investigate how people discover categories for themselves, without labels, feedback, or other forms of external guidance. The problem of discovering new categories, and the learnability issues that it raises, do not arise in studies of conventional discrimination learning.

Testing theories of learning. In our initial experiments using these tasks (Clapper & Bower, 1991), people performed well when shown instances from a single category; that is, they quickly learned which dimensions had consistent

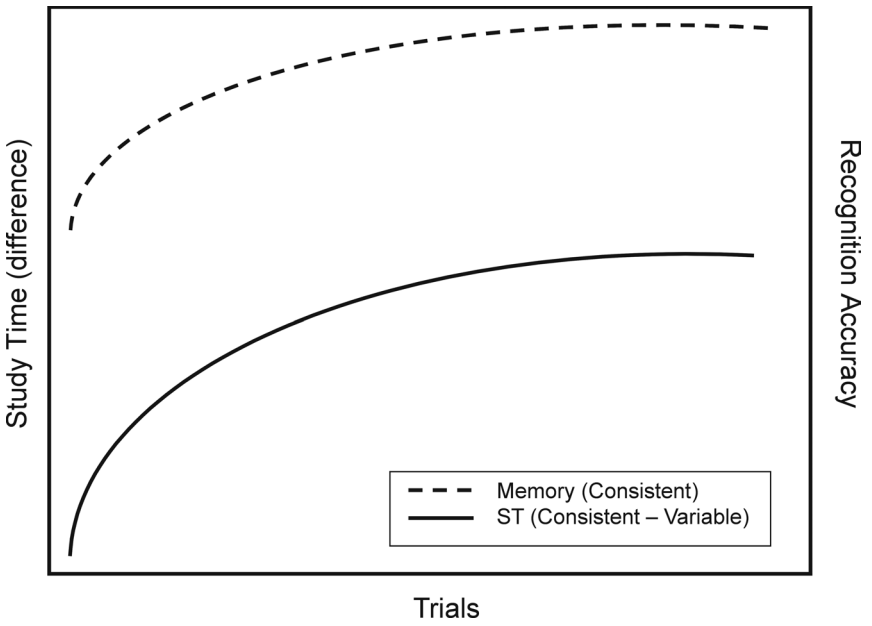

Figure 19–7. Indicators of category learning in the exemplar memory task. As people shift their attention from consistent to variable dimensions, the difference in the amount of study time they receive should gradually increase. Memory for consistent dimensions should also increase over trials.

values within that category, and which were variable. However, when instances of two different categories were shown together in random order, people showed much poorer learning (Clapper, 2006; Clapper & Bower, 1994, 2002). This lack of learning was surprising given the very strong predictive structure that defined the categories in these experiments; most theories would predict that these kinds of categories should be learned very easily.

The simplest model of how people might learn categories in these tasks relies on simple feature association. In such a model, co-occurring features become strongly associated over time and separate feature clusters gradually emerge as the basis for distinct categories. This notion of auto-associative concept learning goes back at least to the mid-18th century (e.g., Bain, 1855/1968; Hume, 1739/2000) and still forms the basis of many learning algorithms today, for example, McClelland and Rumelhart (1985), J. A. Anderson (1977; Anderson, Silverstein, Ritz, & Jones, 1977), Davis (1985), Zeaman and House (1963), and Billman and Heit (1988). Given such a learning process, the kinds of categories used in our tasks should have been very easy for people to learn, contrary to our actual data.

Further research has clarified the conditions under which people will learn categories in these tasks. One factor that has turned out to be very important is

the particular order in which examples of different categories are shown. As already noted, when instances of two categories are presented in a randomly intermixed sequence, relatively little learning of either category is observed. However, when the sequence is arranged so that one category (call it "A") is well-learned prior to the introduction of the second category ("B"), then people will learn both categories easily. Thus, categories A and B will both be learned much better in a "contrast-enhancing" sequence like A **A** A **A A** A **A** A **A A** A A B A B B A B A A A B, and so forth, than in a "mixed" sequence like A **B** A **B B** A **B** A A **B B** A B A B B A B A A A B, and so on. The only difference between these two sequences is the fact that the first six Bs in the second (mixed) sequence are absent and replaced by the same number of As in the first (contrast) sequence. Thus, simply reducing the number of B instances (by eliminating all of them from the first 12 trials of the training sequence) leads to a dramatic improvement in B learning (Clapper & Bower, 1994, 2002). This "negative exposure effect" is a violation of the monotonic (mutually increasing) relationship between exposure and learning normally expected in experiments on human and animal learning.

Interestingly, this effect is incompatible with simple feature association models, all of which are strongly committed to the assumption that additional exposure will lead monotonically to an increase (or at least not a decease) in learning. However, it is easily accommodated within a different type of model, one based on the notion that categories are generated by a discrete, nonincremental "category invention" process (Clapper, 2006; Clapper & Bower, 1994, 2002). This model assumes that people create new categories in response to a specific environmental "trigger" such as a novel or unusual stimulus that fails to match any of a person's existing categories. This kind of learning process implements a "failure-driven" (Schank, 1982) or "win-stay-lose-shift" strategy in which new categories are created only in response to the direct failure of existing categories.

The negative exposure effects described previously are easily accommodated within the category invention framework. To explain poor learning in the mixed condition, it need only be assumed that the first A and B instances appear similar enough for many learners to lump them together into a single category at the start of training. Once this has occurred, they will tend to remain "stuck" in this aggregated state because further A and B instances will continue to fit the combined category; hence, there is never any "surprise event" to trigger the invention of separate categories. By contrast, when people receive a sequence like A A A A A A A A A A A A A B A B B A B A A A B, and so on, they have time to learn category A before seeing any examples of category B. As a result, the first B will be seen as contrasting sharply with previous As, and the learner will invent a new category to accommodate this obvious mismatch. The upshot is that is it easier to add a new category to an already-learned category than to learn two categories at the same time, due to the potential for aggregation in the latter situation.

An interesting characteristic of category invention models is that they posit a dichotomous-state conception of learning: A learner generates separate cate-

a) Opposite Themes	b) Neutral (No Themes)	c) Same Theme
Category A favorite drink: **sherry** favorite activity: **yachting** enjoys watching: **the opera** drives: **a Mercedes** clothes by: variable lives in: Sacramento employed as: **a lawyer** TV show: variable favorite food: steak last vacation: New Orleans favorite music variable graduated: **Harvard**	**Category A** favorite drink: coffee favorite activity: soccer enjoys watching: basketball drives: a Honda clothes by: variable lives in: Sacramento employed as: an accountant TV show: variable favorite food: steak last vacation: New Orleans favorite music by: variable graduated: community college	**Category A** favorite drink: **sherry** favorite activity: **yachting** enjoys watching: **the opera** drives: **a Mercedes** clothes by: variable lives in: Sacramento employed as: **a lawyer** TV show: variable favorite food: steak last vacation: New Orleans favorite music variable graduated: **Harvard**
Category B favorite drink: **beer** favorite activity: **bowling** enjoys watching: **pro wrestling** drives: **an old Pickup** clothes by: variable lives in: Tucson employed as: **is unemployed** TV show: variable favorite food: fish last vacation: San Francisco favorite music by: variable graduated: **high school**	**Category B** favorite drink: cola favorite activity: softball enjoys watching: movies drives: a Toyota clothes by: variable lives in: Tucson employed as: a technician TV show: variable favorite food: fish last vacation: San Francisco favorite music by: variable graduated : state university	**Category B** favorite drink: **fine wine** favorite activity: **polo** enjoys watching: **the symphony** drives: **a Lamborghini** clothes by: variable lives in: Tucson employed as: **a plastic surgeon** TV show: variable favorite food: fish last vacation: San Francisco favorite music by: variable graduated: **Princeton**

Figure 19–8. Sample categories used in experiments on prior knowledge and schema induction. Thematic consistent values within each category are indicated by boldface type, neutral values by plain type, and variable values are underlined (all were shown in plain type in the actual experiments).

gories for different subsets of stimuli, whereas a nonlearner fails to do so and aggregates all the stimuli into a single category. In principle, this overaggregated state should be very stable over time. Clapper (2006) provides an example in which initial aggregation prevents people from later acquiring separate categories from a sequence in which they otherwise would easily do so. In this experiment, categories A and B were learned more poorly in sequence **A B A B B A A B** A A A A A A A A A A A B A B B A B A A A B, and so on, than in A A A A A A A A A A A B A B B A B A A A B, and so on. However, the only difference between these sequences is that several instances of both categories have been added to the beginning of the second sequence in order to create the first sequence. Thus, simply adding a few instances of both categories to an already "good" training sequence had the paradoxical effect of turning it into a "bad" sequence, a dramatic violation of monotonicity. Once the categories became aggregated during the early mixed trials, people were apparently unable to "de-aggregate" them during the trials that followed. (Additional experiments have shown that if the interval of all-A trials is made longer, say, 24 instead of 12 instances, people are able to de-aggregate and learn the two separate categories.)

Optimality issues. While these otherwise paradoxical results make sense within the category invention framework, they also seem to contradict commonsense intuitions about rational design: The insensitivity to category structure and extreme order sensitivity demonstrated in these experiments seem to suggest a highly nonoptimal learning process. Surprisingly, this turns out not to be the case: A normatively optimal category invention algorithm *can* produce the same patterns of sequence sensitivity observed in our experiments, given certain reasonable assumptions about memory and learning biases.

Our explorations of this issue (Clapper, 2006; Clapper & Bower, 2002) have been carried out using a Bayesian ideal-observer model proposed by J. R. Anderson (1990, 1991). This model, known as the rational model of categorization, provides an example of a formally specified computational theory that fits our qualitative category invention framework. In its usual mode of application, the model assumes perfect memory for each training instance. In this mode, it is highly sensitive to the correlational structure of the input stimuli regardless of training sequence—which means that it cannot simulate the results described earlier. However, if the model is "weakened" by reducing memory for individual training instances and by assuming a low a priori estimate of the likelihood of new categories, very different results can be obtained. Under these conditions, the model becomes highly sensitive to training sequence and can reproduce the nonmonotonic sequence effects (negative exposure effects) described previously.

What makes this supposedly optimal algorithm so sensitive to training sequence under certain parameter settings? The answer to this question lies partly in the structure of the model itself, and partly in human processing limitations that can be simulated within a specific region of the model's parameter space. The rational model is a dichotomous-state model that implements a win-stay-lose-shift learning procedure. If it happens to assign the first B in a mixed sequence to the same category as the previous A(s), then it will find itself in an aggregated state from which it will usually be unable to recover. This will not occur if a long sequence of A's is presented before any B's, or if only one or two A's have been shown but the model remembers them accurately enough to avoid lumping them with later B's.

Of course, poor memory would have completely different effects on a different type of learning process. For example, in a feature association model poor memory might translate into a lower-than-usual learning rate, but this would not alter the fundamentally monotonic nature of the model itself—more instances would still result in better learning, all else being equal.

CURRENT DIRECTIONS

Knowledge effects. The previous experiments investigated how people use observed training instances to generate and learn new categories. But empirical observation is not the only foundation on which categories are built; a large body

of research shows that prior knowledge and intuitive theories also play an important role (e.g., Heit & Bott, 2000; Murphy, 2002; Murphy & Medin, 1985).

I have carried out several experiments (Clapper, 2005, in press) on these issues using the exemplar memory task and categories related to familiar themes (i.e., personality stereotypes such as highbrows vs. lowbrows, young vs. old, male vs. female). It is clear that prior knowledge (thematic relatedness) helps people discover separate categories in this task. Thus, people were much more likely to notice that the training instances could be divided into two categories if those categories were related to different themes, e.g., highbrows vs. lowbrows (Fig. 19–8a) as opposed to "average" or "normal" individuals (Fig. 19–8b). The themes also helped people learn the features within each category. In particular, features that were related to the themes (e.g., enjoys watching the opera vs. pro wrestling for the highbrow/lowbrow themes) were learned more quickly than "neutral" features that were unrelated to the themes (e.g., lives in Sacramento vs. Tucson; note the two different types of consistent features within the thematic categories in Figures 19–8a and 19–8c).

It is possible to test hypotheses concerning exactly *how* knowledge facilitates the learning of thematic features in this task. One possibility is a *within*-category effect in which knowledge binds and integrates the features within a category, thereby making them easier to learn and remember. Another possibility is a *between*-category effect in which knowledge reduces the amount of confusion and interference that would otherwise occur between the features of related categories. A within-category effect should be detectable whether a category is presented alone or with another category; also, in a two-category task, it should not matter whether the categories are related to the same theme or different themes, just so long as they are related to *some* theme that can help bind their features together. A between-category effect, by contrast, cannot occur unless more than one category is present and only when those categories are related to different themes (so that the themes can provide a basis for telling their features apart). Several experiments (described in Clapper, 2005, in press) have provided clear evidence for between-category effects—in other words, less interference between the features of related categories when those categories are related to contrasting themes, e.g., highbrows vs. lowbrows. However, these experiments showed no evidence for within-category effects—in other words, no benefit of thematic relatedness in a single-category task, and no benefit in a two-category task when both categories were related to the same theme (as in Fig. 19–8c). These results are especially interesting because previous research has tended to stress the importance of knowledge in promoting within-category coherence (e.g., Kaplan & Murphy, 1999; Murphy & Kaplan, 2000; Spalding & Murphy, 1996), whereas the possibility of between-category knowledge effects has been almost completely ignored.

The structural basis of categories. The idea of interference across category boundaries raises some interesting questions about how category schemas are actually represented in memory. It is often convenient to think of schemas as though they are isolated, independent data structures, each existing separately

from the larger web of knowledge. But that cannot be correct. In fact, what I am calling schemas may not exist as unitary, precomputed structures at all, but might actually consist of many smaller components distributed throughout memory, with each component shared by many categories (Rogers & McClelland, 2004). At a minimum, schemas must interact with other schemas in order to capture the generative capacity of human reasoning (McClelland, Rumelhart, & Hinton, 1986). In general, it seems reasonable to assume that schemas representing related categories might contain overlapping representational structure, and that the properties of one schema might affect the learning or the stability of other schemas. If so, then one way to learn more about how schemas are represented might be to investigate how they affect (e.g., interfere with) other schemas.

One way to study interference effects is to manipulate the degree of structural overlap between different categories and observe the effects on learning within each category. The stimulus sets used in the experiments described so far—and in the vast majority of category-learning experiments ever conducted—contained stimuli that all varied along the same attribute dimensions, with categories defined in terms of specific values or clusters of values along these shared dimensions. In the terminology of Garner (1974), stimuli within these sets shared a common *dimensional* structure whereas different categories were defined in terms of *correlational* structure. In the experiments discussed next, people's learning of categories based on correlational structure was directly compared with that of categories based on dimensional structure (Fig. 19–9); that is, categories that varied along the *same* dimensions (Fig. 19–9a) were compared to categories that varied along *different* dimensions (Fig. 19–9b). In the language of structure-mapping theory (e.g., Gentner, 1983; Gentner & Markman, 1997), this is a contrast between categories with *alignable* dimensions vs. categories with *non-alignable* dimensions.

The main result of this research is that correlation-based categories suffer considerable interference along their shared dimensions (Clapper, 2004). I have already described how people have difficulty discovering separate categories in these kinds of stimulus sets. Further experiments have shown evidence of feature-level interference even after it is clear that separate categories have already been recognized. For example, even if people have already learned two categories in the exemplar memory task, presenting a third category has a strongly negative effect; not only is the third category learned poorly compared to the previous two, but levels of learning for the previous categories also decline following the introduction of the third category (Clapper, 2004). However, such interference is observed only when the third category has the same dimensional structure as the first two. In general, when the categories in a set are distinguished at the level of dimensional structure, people perform as though they were learning each category in isolation (i.e., in a one-category task). They immediately partition the set into separate categories, show no interference in learning the features within each category, and learning of each category is unaffected by how many other categories are present.

Research on interference and other schema-level interactions may provide useful information about how category knowledge is represented in memory; it may

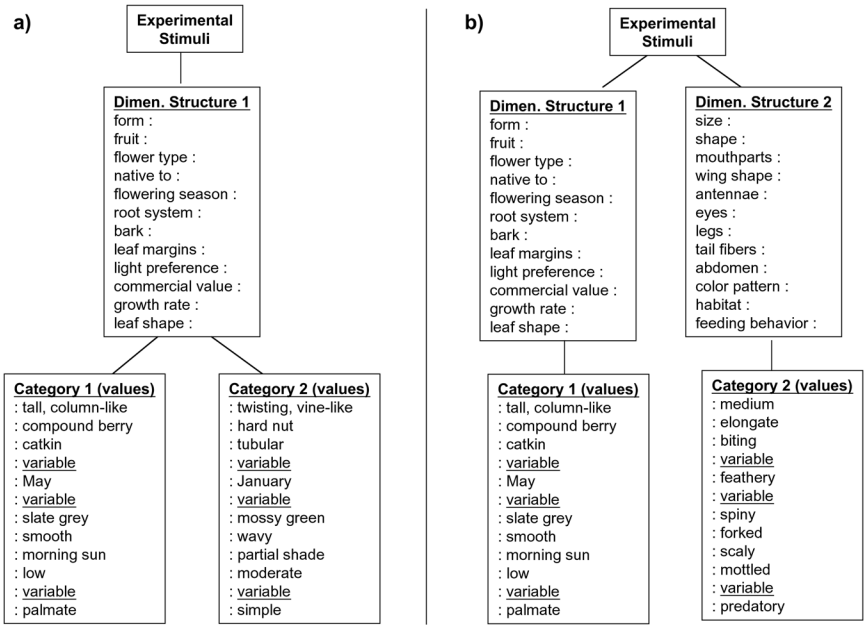

Figure 19–9. Sample categories defined by (a) correlational vs. (b) dimensional structure. Note that both categories in (a) vary along the same dimensions, while the two categories in (b) vary along different dimensions.

also serve as an antidote for the simplifying tendency to think of schemas as isolated, encapsulated "boxes in the head." In fact, the current paradigm may provide a context in which many hypotheses related to schema structure and schema-level interactions can usefully be tested.

SUMMARY AND CONCLUSIONS

This chapter began by discussing the contrast between discrimination learning and schema induction as frameworks or "paradigms" for the investigation of category learning. I argued that discrimination learning fails to capture some important aspects of category learning and that a number of issues that have been poorly articulated or difficult to study within the discrimination-learning framework happen to be much more approachable within a schema induction framework.

One of these issues concerns how category knowledge affects instance processing and memory organization when instance learning, rather than categorization, is the proximal task goal (as assumed by schema induction). This issue is of major importance to researchers who study the role of schemas in problem solving, reasoning, language comprehension, and other higher level cognitive

abilities. However, for the most part they have studied this issue using naturally occurring schemas acquired outside the laboratory, and that research has had little direct connection with studies of artificial category learning. The first experiment described earlier (and several others described in Clapper & Bower, 1991) attempts to bridge the gap between research on artificial category learning and studies of schema application. The greater control enabled by synthesizing knowledge structures to our own specifications allowed us to test hypotheses about memory organization that would have been difficult or impossible to test with natural materials.

A second issue concerns what might be called the "discovery problem" in category learning—namely, the issue of how people detect the existence of separate new categories without externally provided labels or corrective feedback to guide them. Obviously, this issue does not arise in supervised classification tasks. Several experiments (Clapper, 2006; Clapper & Bower, 1994, 2002) were described that provided strong evidence that this process conforms to a category invention framework first described in Clapper and Bower (1991) and elaborated in Clapper and Bower (1994, 2002) and Clapper (2006).

More recent experiments have focused on the role of prior knowledge in category learning, specifically how prior knowledge is used to help people discover new categories and the role of such knowledge in helping to reduce feature-level interference between closely related categories. Further experiments have focused on those between-category interference effects and have shown that they depend critically on the existence of shared dimensional structure between the categories, along with specific assumptions about memory and task biases. All of these experiments deal with issues that have received little attention by researchers working within the standard discrimination-learning paradigm.

I'd like to conclude this chapter with a word of thanks to Gordon, who was there from the beginning and who has supported and contributed to this research in so many ways. Gordon always had a knack for encouraging creative thinking and new approaches among his students and co-workers, while at the same time adhering to the highest standards of rigor and excellence. I hope this research has proved worthy of those high standards, and that it will continue to generate new findings and new surprises in the years ahead.

REFERENCES

Anderson, J. A. (1977). Neural models with cognitive implications. In D. LaBerge & S. J. Samuels (Eds.), *Basic processes in reading: Perception and comprehension* (pp. 27–90). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Anderson, J. R. (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.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Anderson, J. R., & Bower, G. H. (1974). Interference in memory for multiple contexts. *Memory & Cognition, 2,* 509–514.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review, 61,* 183–193.

Bain, J. (1868). *The senses and the intellect*. New York: Appleton. (Original work published 1855).

Bartlett, F. C. (1932). *Remembering: A study in social psychology*. Cambridge, England: Cambridge University Press.

Bellezza, F. S., & Bower, G. H. (1981). The representation and processing characteristics of scripts. *Bulletin of the Psychonomic Society, 18,* 1–4.

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

Bobrow, D. G., & Norman, D. A. (1975). Some principles of memory schemata. In D. G. Bobrow & A. Collins (Eds.), *Representation and understanding.* New York: Academic Press.

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

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

Chin-Parker, S., & Ross, B. H. (2004). Diagnosticity and prototypicality in category learning: A comparison of inference learning and classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 216–226.

Clapper, J. P. (2004, November). *Dimensional structure, alignability, and unsupervised learning*. Poster presented at the annual convention of the Psychonomic Society, Minneapolis, MN.

Clapper, J. P. (2005, November). *Within- vs. between-category knowledge effects in observational learning*. Poster presented at the Annual Convention of the Psychonomic Society, Toronto, Ontario, Canada.

Clapper, J. P. (in press). Prior knowledge and correlational structure in unsupervised learning. *Canadian Journal of Experimental Psychology*.

Clapper, J. P. (2006). When more is less: Negative exposure effects in unsupervised learning. *Memory & Cognition, 34*, 890–902.

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, pp. 65–108). New York: Academic Press.

Clapper, J. P., & Bower, G. H. (1994). Category invention in unsupervised learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 443–460.

Clapper, J. P., & Bower, G. H. (2002). Adaptive categorization in unsupervised learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 908–923.

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

Garner, W. R. (1974). *The processing of information and structure*. New York: Lawrence Erlbaum Associates.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 155–170.

Gentner, D. & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist, 52*, 45–56.

Graesser, A. C., (1981). *Prose comprehension beyond the word*. New York: Springer-Verlag.

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.

Heit, E., & Bott, L. (2000). Knowledge selection in category learning. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 39, pp. 163–199). San Diego, CA: Academic Press.

Hume, D. A. (2000). *A treatise of human nature*. New York: Oxford University Press (Original work published 1739).

Kaplan, A. S., & Murphy, G. L. (1999). The acquisition of category structure in unsupervised learning. *Memory & Cognition, 27,* 699–712.

Kling, J. W., & Riggs, L. A. (1971). *Experimental psychology* (3rd ed.). New York: Holt, Rinehart & Winston.

Markman, A. B., & Ross, B. H. (2003). Category use and category use and category learning. *Psychological Bulletin, 129*, 592–613.

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

McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986). The appeal of parallel distributed processing. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 3–44). Cambridge, MA: MIT Press.

Medin, D. L., & Schaffer, M. M. (1978). A context theory of classification learning. *Psychological Review, 85,* 207–238.

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* (pp. 331–364). Palo Alto, CA: Tioga.

Minsky, M. (1975). A framework for representing knowledge. In P. H. Winston (Ed.), *The psychology of computer vision* (pp. 211–277). New York: McGraw-Hill.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Murphy, G. L., & Kaplan, A. S. (2000). Feature distribution and background knowledge in category learning. *The Quarterly Journal of Experimental Psyc*hology, *53A*, 962–982.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289–316.

Oldfield, R. C. (1954). Memory mechanisms and the theory of schemata. *British Journal of Psychology, 45,* 14–23.

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

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology, 83,* 304–308.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach.* Cambridge, MA: MIT Press.

Rumelhart, D. E., & Ortony, A. (1977). The representation of knowledge in memory. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), *Schooling and the acquisition of knowledge* (pp. 99–135). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

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

Spalding, T. L., & Murphy, G. L. (1996). Effects of background knowledge on category construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 525–538.

Trabasso, T., & Bower, G. H. (1968). *Attention in learning: Theory and research*. New York: Wiley.

Zeaman, D., & House, B. J. (1963). The role of attention in retardate discrimination learning. In N. R. Ellis (Ed.), *Handbook of mental deficiency* (pp. 159–223). New York: McGraw-Hill.