Psychologica Belgica
1995, 35-2/3, 127-144.

KNOWLEDGE, SIMILARITY, AND CONCEPT FORMATION

Gregory L. MURPHY
University of Illinois, USA

Thomas L. SPALDING
University of Iowa, USA

This article briefly reviews two recent lines of research that investigate the role of background knowledge in category learning and category formation. The results of several experiments show that when subjects are able to use their background knowledge to integrate the features of instances with respect to an underlying theme, they learn the categories being taught by the experimenter much more easily. In addition, the opportunity to apply background knowledge dramatically affects what categories the subjects spontaneously form when they do not receive feedback from an experimenter. The implications of these results for traditional accounts of similarity in category learning and formation are discussed.

People's concepts generally group together similar things. The categories that are formed in natural language, such as *bird*, *table*, *bowl*, *computer* or *lawyer*, group together objects (or people) that have a number of characteristics in common. This observation has led many theorists of the psychology of concepts to draw a causal connection between the similarity of the instances and the formation of the concepts. As is well known, early researchers on this topic (e.g., Hull, 1920) assumed that there was a small set of characteristics that were distinctive to the members of the category. This is what has become known as the *classical theory* of concepts. More recently, researchers have largely abandoned that claim as being overly strong (see Smith & Medin, 1981) and have instead argued for the weaker position that there are characteristics that are generally common to members of the category and not generally common to members of other categories. Thus, the members of the category are all similar to one another and usually less similar to nonmembers (see, e.g., Homa, Rhoads, & Chambliss, 1979; Posner & Keele, 1970; Rosch & Mervis, 1975). That is, there is something of a *Family Resemblance* structure to concepts.

As a description of category structure, the view that concepts include similar things is probably correct in most cases. However, as an explanatory mechanism, this claim about similarity leaves much to be desired. We will just mention a few of the problems, which have been described at length elsewhere (e.g., Medin, 1989; Murphy, 1993b; Murphy & Medin, 1985). One problem is that

---

This research was supported by NIMH grant MH41704. We would like to thank Stephanie Doane, Lisa Oakes and Jodie Plumert for helpful comments on this paper.

Correspondence concerning this article should be addressed to Gregory L. Murphy, Beckman Institute, University of Illinois, 405 No. Mathews Avenue, Urbana, IL 61801, USA. Electronic mail may be sent to gmurphy@s.psych.uiuc.edu.

of cross-categorization. Many objects are members of a number of different categories: *dog, pet, animal, carnivore, guard dog*; or *apple, fruit, food, toy, present*; or *mother, scientist, cook, teacher, conservative*. If objects are grouped together according to their similarity, it is surprising that the same object should appear in so many different categories. Why is it that Molly (a dog) is similar to all other dogs in one case, but similar to all the pets in another case? This is especially puzzling when we consider that some dogs are not pets, and so are similar to Molly in one case but not the other. Another problem is that the features that are used to establish similarity can be different for different categories. For example, Molly is similar to other dogs because they (almost) all have fur, four legs, bark, etc. However, Molly is similar to other pets because they (almost) all are friendly, cuddly, tame, live in the home, etc. Why is it that "tame" is a relevant feature for thinking about Molly as a pet but not as a dog?

The fact that different features are used in computing similarity is a real problem, because the similarity between items can be dramatically different, depending on what features are used. Thus, Molly is similar to Sierra, a wild dog, by virtue of sharing some biologically relevant features. But if we were to specify other features as being equally important, such as where Molly lives, her relations with people, her grooming, her training, etc., Molly is no longer very similar to Sierra. She might be more similar to Muppet, a cat. The point is that similarity does not seem to be a primitive that can be computed independently of a way of thinking about the object. And that way of thinking about the object is very close to the concept itself. That is, when we think about Molly as a dog (i.e., categorize her), some features and a certain way of calculating similarity become relevant; when we think about Molly as a pet (i.e., place her in a different category), other features and a different similarity metric become relevant. Thus, there is a threat of circularity in using similarity as a theoretical explanation for categories, because similarity may presuppose a categorization of the object (see also Medin, Goldstone, & Gentner, 1993).

As a result of such problems, we need to consider other possible determinants of conceptual structure that may help break us out of this circularity. We consider here the contribution of "knowledge" of a domain. This knowledge includes deep, underlying theories about the domain (such as the genetic basis of an animal's characteristics), as well as specific facts. We will argue that this knowledge has an important influence on the formation and learning of new concepts. Typically, new concepts are learned in a domain that we already know something about. We already know many kinds of plants and animals, yet we may encounter a new kind in the zoo or arboretum; we may already know about computers or cars, yet new kinds appear quite regularly; we know a fair amount about people and their professions, but we may encounter a new type of personality or a new profession. The point is that even when we learn new concepts, it is usually in the context of a large amount of related information

that we already know. This knowledge has powerful effects on the way that we learn new concepts.

Much research on concept learning has ignored the effect of knowledge. In fact, experiments are typically designed to avoid any knowledge influences, by using stimuli that are novel and are associated with no domain. For example, colored geometric figures (circles, triangles, etc.), dot patterns, color patches, schematic faces, alphanumeric sequences and other artificial stimuli have been popular stimulus materials in this field. Even more realistic stimuli, such as novel animals, are often chosen so as to be clearly fictional in order to avoid the knowledge that one might otherwise bring to them (e.g., they may be described as coming from an unknown planet). This kind of research may well provide important information about how people learn concepts. It is certainly true that one cannot learn concepts based solely on one's knowledge or prior expectations about a category—the empirical evidence from category members must play a role as well. However, research that is specifically designed to remove the possibility of using knowledge cannot possibly tell us anything about how knowledge is involved in learning and forming more natural categories. Furthermore, such purely empirical approaches may well be misleading, in that their assumptions may not be true when learners have knowledge about a domain. That is, some have argued that it is impossible to simply "add on" the knowledge to a purely empirical concept-learning model—the empirical and knowledge components must be closely integrated (e.g., Wisniewski, 1995; Wisniewski & Medin, 1994).

We are not going to criticize these purely empirical approaches here (see Murphy, 1993a, 1993b, for a longer discussion), though we shall return to them in the final section of the article. Instead, we want to focus on the things that such approaches do not address, namely the ways that knowledge might be involved in concept formation and learning. One important use of knowledge is in specifying the relevant features in a domain. For example, our knowledge of biology tells us to look for genetic and structural attributes in learning a new kind of plant. In contrast, our knowledge of professions tells us to look for the activities, education, and employer of the person in understanding a new profession. Our knowledge of these domains helps to focus our attention on certain attributes, and it may even partially define what those attributes are (Wisniewski & Medin, 1994).

Another possible influence of knowledge is in helping the learner to tie together the attributes in the concept once they are identified. For example, a tiger is extremely fast, has sharp claws, large teeth, and hunts animals for food. Our knowledge of predators and their typical behavior helps to provide a way of tying together these features: A predator must be fast in order to catch its prey, and it must have sharp teeth and claws in order to kill and eat it. It seems likely that this kind of explanation of the features makes it easier for people to

learn the concept of *tiger* than if we did not have the knowledge to connect the features. It is this use of knowledge that we will explore further in this article.

Concept Learning

The use of knowledge in concept learning has been documented fairly extensively. Pazzani (1991) examined the influence of knowledge on learning conjunctive and disjunctive rules that defined a category. For example, a conjunctive rule might be that members of the category are red AND square, and a disjunctive rule would be that they are red OR square. Although people normally learn conjunctive rules more easily, they learned the disjunctive rule more easily when it was particularly related to knowledge about the category. Similarly, Wattenmaker, Dewey, T. Murphy, and Medin (1986) showed that different formal category structures could not in and of themselves predict learning rates of categories. They discovered that when a given category structure was consistent with knowledge that could be brought to bear on it, learning was easy. When the category structure and knowledge were inconsistent, learning was hard. The category structure itself was not a sufficient predictor of learning.

In a recent set of studies, Murphy and Allopenna (1994) investigated two possible ways that knowledge might affect concept learning. One idea was that the features of natural concepts are usually more familiar and semantically richer than the features of artificial concepts. As already mentioned, most studies of concept learning use impoverished features such as geometric shapes, dots, color patches, and the like. However, in real objects, the features are often more complex and more interesting. A second idea about how knowledge might influence concept learning has to do with the relations among the features. If an experiment uses simple perceptual features, such as geometric shape, length, color, number, etc. to define the stimuli, then there will generally not be any interesting knowledge relating the features: There is no reason to expect that a circle should be blue instead of red. However, when more meaningful features are used, then some relations can be identified. The relations among some features give rise to expectations about other possible features. For example, if a person is "friendly" and "outgoing," then you might expect that she would "like to go to parties" but would not be described as "moody." The knowledge that one feature evokes might be consistent or inconsistent with knowledge evoked by the other features. As a result, the knowledge could influence the learning of features associated with a category.

In their first experiment, Murphy and Allopenna compared three different kinds of categories, which varied in these two aspects of world knowledge (meaningfulness of individual features and integration of features). The first

category type, called the *Arbitrary Condition*, used relatively meaningless and semantically impoverished features, typographical characters such as parentheses, the dollar sign, and the equal sign. Two representative examples are shown in Table 1 (which we explain in more detail below). Another condition, the *Meaningful Condition*, used meaningful English phrases, like "Lives alone" and "Modern furniture." However, these phrases were grouped together in categories so that they did not have any apparent connection. For example, something that "Lives alone" might also be "Made in Africa" and have "Thick, heavy walls" (see Table 2). There is no simple knowledge structure that connects these features. The final condition was called the *Integrated Condition*. Here, there was a theme that united the features. For example, the building categories in Table 3 show a clear theme related to the location and structure of the buildings. One category describes a kind of underwater building, and the other describes a building that floats in the air. The concept's features are related to this theme.

Table 1

Structure of Arbitrary Categories

Features Regularly Associated with Each Category	
Category 1	*Category* 2
+	-
{	}
>	<
$	@
[	]
Examples of Each Category	
Category 1	*Category* 2
?	?
+	-
[	]
=	"
>	<
>	=
!	]
{	?
$	@
"	<

Note. Each example was printed on a separate card, with each feature on a different line, much like the examples shown here.

Table 2

Structure of Meaningful Categories

Features Regularly Associated with Each Category	
Category 1	*Category 2*
Lives alone	Lives in groups
Made in Africa	Made in Norway
Fish kept there as pets	Birds kept there as pets
Thick heavy walls	Thin light walls
Has barbed tail	Has furry tail
Examples of Each Category	
Category 1	*Category 2*
Lives alone	Lives in groups
Four door	Doesn't hibernate
Modern furniture	Modern furniture
Thick heavy walls	Has furry tail
Has barbed tail	Birds kept there as pets
Two door	Victorian furniture
Fish kept there as pets	Lives in groups
Hibernates	Made in Norway
Has barbed tail	Doesn't hibernate
Thick heavy walls	Birds kept there as pets

Note. The features used in this condition were the same as those used in the Integrated condition, but mixed up so that features from different domains were now in the same category.

Each category was associated with five features (listed under it at the top of each table). Individual learning examples were constructed by selecting a subset of these "correct" features and combining them with nonpredictive features. Thus, no feature was necessary for classification, and the categories had considerable overlap, because of the nonpredictive features that were found in exemplars of both categories. In order to learn the categories, subjects had to learn which features were reliably associated with each category. This task was formally the same for all conditions. That is, each category pair had the same number of features, and learning examples were constructed from the features in exactly the same way. All that varied across conditions was the content of the features themselves. The Arbitrary Condition used meaningless features and allowed no integration, the Meaningful Condition used meaningful features (in fact, they were the same features as in the Integrated condition), but also allowed no integration, and the Integrated Condition allowed the use of both kinds of knowledge: The individual features were meaningful, and the features could be integrated. By comparing these three conditions, Murphy and Allopenna hoped to discover whether either kind of knowledge affected concept learning.

Subjects attempted to learn the categories by viewing examples like those

Table 3
Structure of Integrated Categories

Features Regularly Associated with Each Category	
Category 1	*Category 2*
Fish kept there as pets	Birds kept there as pets
Thick heavy walls	Thin light walls
Divers live there	Astronauts live there
Under the water	Floats in the air
Get there by submarine	Get there by plane
Examples of Each Category	
Category 1	*Category 2*
Fish kept there as pets	Astronauts live there
12-month lease	Floats in air
Has rugs	6-month lease
Divers live there	Thin light walls
Get there by submarine	Has wall-to-wall carpeting
Modern furniture	Has rugs
Under the water	Get there by plane
Thick heavy walls	Victorian furniture
Get there by submarine	Astronauts live there
6-month lease	Birds are kept there as pets

Note. There were three category pairs in this condition: buildings (shown), animals and vehicles.

shown in the bottoms of Tables 1-3. They were required to guess which category each example was in, and they cycled through blocks of 20 examples until they classified all items correctly. Feedback was given after each guess.

Results showed that subjects learned the Integrated categories faster than either the Meaningful or Arbitrary categories, which were equally difficult to learn (see Murphy & Allopenna, 1994, for a detailed description). Thus, these results suggest that the beneficial effect of knowledge is not due to the semantic richness or meaningfulness of the features per se, but rather to the way that knowledge ties together the features of a concept. That is, knowledge about individual features was not particularly helpful, but knowledge about feature *relations* was extremely helpful.

One potential problem is that the Meaningful Condition's features are actually contradictory. For example, if something is "Made in Africa," then how can it have the feature "Lives alone"? One feature seems to be appropriate for an artifact, and the other for an animal. This contradictory aspect of the features arose from a desire to have the same features in the Meaningful and Integrated Conditions. When mixing up the features of the Integrated Condition, such contradictions inevitably arise. Perhaps, though, the observed difficulty in learning the Meaningful categories was the result of such (unnatural) contradictions, rather than a more positive use of knowledge to help the

Integrated categories. (Note, however, that the features are only contradictory with respect to the knowledge one has about the kinds of properties that are true of artifacts and animals.) To investigate this hypothesis, Murphy and Allopenna constructed a different kind of category, called the *Domain-Consistent Condition*. Domain-Consistent features all come from a single domain (e.g., a kind of building or a kind of animal), but cannot be integrated by a theme. For example, one kind of building had features such as "central heating, brick exterior, patio, large front yard," whereas its contrast category had "non-central heating, wood exterior, porch, large back yard." The features of each category are fully consistent—there are no contradictions. At the same time, however, there are no knowledge-based connections within the category. That is, there is no particular reason to expect that a house with a brick exterior would have a large front yard, and a house with a wood exterior would have a large back yard. Thus, the individual features are meaningful, and do not contradict each other, but are unlikely to be integrated. In this experiment, Murphy and Allopenna compared the Integrated Condition to the Domain-Consistent and Meaningful Conditions.

In fact, the Meaningful and Domain-Consistent categories were about equally difficult to learn, whereas the Integrated categories were significantly easier. The Integrated categories were learned in 2.2 blocks, which is extremely fast given that 2 is the effective smallest number of blocks in which the categories could be learned (in the first block, subjects must guess on the first few stimuli, and perfect performance on a block was required for learning). The Meaningful categories took 5.2 blocks to learn and the Domain-Consistent categories 4.1 (this difference was not significant). In test items given after learning was complete, the Domain-Consistent condition was no more accurate than the Meaningful group, but the Integrated subjects were faster and more accurate than either. Thus, it appears that having features that simply do not contradict one another is not very helpful in concept learning. Rather, the more positive relations between the features in the Integrated condition seem to be important.

Murphy and Allopenna (1994) suggested that the utility of knowledge in these kinds of categories must lie in the way that it helps one to predict what features a category will have. For example, once you know that a building is underwater, you can predict with relative certainty whether it will have thick or thin walls. This means that you do not need to go through a long, onerous process of learning associations of the features (or exemplars) to each category. One implication of this conclusion is that when knowledge is used in this way, subjects may not learn the empirical distribution of individual features very well. Murphy and Allopenna found some evidence that subjects were not as sensitive to the frequency of individual features when there was such knowledge. In the experiment just described, some of the features were presented frequently,

and others occurred only once per block. After learning, subjects were required to say which category every feature was associated to, and they also gave a typicality rating. Subjects in the Meaningful and Domain-Consistent conditions were much more accurate in classifying the frequent features than the infrequent features. They rated the infrequent features as being quite atypical. Subjects in the Integrated condition showed no such difference: They were very accurate in classifying even features that had appeared only twice, and they did not rate the infrequent features as very atypical. In short, the subjects in the Integrated Condition were relatively insensitive to frequency differences among the features. Thus, when knowledge structures can be easily constructed in order to classify objects, subjects may not learn as much empirical information about individual features or exemplars.

Category Formation

In a typical category learning experiment, the experimenter decides a priori what the categories are, and there is some sort of teaching episode in which subjects try to learn the category that the experimenter has defined. After each trial, the experimenter (or computer) tells the subject what the correct answer is. However, there are examples in real life in which people make up their own concepts, without external feedback. For example, it seems likely that children have learned many natural-language categories before they learn the names for those categories (e.g., Clark, 1981; Merriman, Schuster, & Hager, 1991; Mervis, 1987). Imagine that you are taking a vacation in a foreign country. You may begin to notice that the trees are not the same as the ones you are used to seeing. In fact, there is one particular type that seems rather distinctive. In order to make this observation, you had to notice the similarity of a class of trees, to form a new concept that distinguishes it from the other classes of trees that you know about. In doing so, you noticed the similarities among these new trees (in shape, size, coloring, leaf type, etc.), and ignored some differences (in location, age, etc.). Furthermore, you did all this without a teacher saying "That is a beech... that is an elm... that is an oak... there's another beech."

Since, as we noted earlier, objects can be seen as similar in a very large number of respects and dissimilar in just as many respects, you probably used your knowledge of trees to decide which features to pay attention to, and which to ignore, in noticing this class of trees. You probably were also able to discount some differences, such as size differences when the age of the tree could explain the difference; or differences in foliage when the season has changed. In several experiments, we have begun to investigate how it is that knowledge may influence people's ability to notice categories without any external feedback (Spalding & Murphy, in press).

Past work on this topic has shown that in experimental contexts, subjects are extremely likely to create categories based on a single dimension (Ahn & Medin, 1992; Medin, Wattenmaker, & Hampson, 1987). For example, if subjects are given a set of cards, each describing an animal, they might divide the animals into those with gray vs. brown fur, even if these items could be sorted into two Family Resemblance categories. The result is that all members of the category are identical on that one dimension, but they are overly diverse on all other dimensions; thus, the members of each category are not very similar. In an extensive series of studies, Medin et al. (1987) showed that this tendency to sort unidimensionally was robust to changes of materials and instructions. This result is very puzzling, because most natural categories are not unidimensional—they contain clusters of correlated attributes (Rosch & Mervis, 1975).

In our studies, we have been investigating whether using categories that are unified by a theme, like the Integrated categories described above, would override the unidimensional bias and allow subjects to discover the categorical structure underlying the creation of the exemplars (see also Ahn, 1990, 1991). We gave people exemplars like those shown in Tables 2 and 3 for the Integrated and Meaningful categories, 10 exemplars from each category. We asked them to read through the cards and then told them that the cards actually represented "two kinds of things." They were asked to divide the cards into the two kinds. For the Meaningful stimuli, only 17% of the subjects recovered the category structure, and 58% sorted unidimensionally; for the Integrated stimuli, 75% successfully identified the categories, and only 8% sorted unidimensionally. (The number of unidimensional sorts in the Meaningful group is a bit lower than one would expect from Medin et al.'s results, but this is due to differences in the category structure used—see Spalding & Murphy, in press, for details.) In short, it seems that when knowledge helps to connect the features, subjects spontaneously notice the categorical structure, producing sorts that preserve the family resemblance structure of the categories. In the absence of such knowledge, the category is seldom noticed, and unidimensional sorts predominate.

As mentioned earlier, the Meaningful stimuli include contradictory features, which might have kept subjects from noticing the categorical structure. It may not be surprising that subjects don't pay attention to all features when they are as inconsistent as "Modern furniture" and "Has a bushy tail." Thus, we also tested a Domain-Consistent condition, comparing it to the Integrated group. In this experiment, we also made some changes in the procedure to make it more closely match that of Medin et al. (1987). The percentage of subjects who discovered the category structure was 6% in the Domain-Consistent condition and 44% in the Integrated condition, a significant difference. These numbers are somewhat lower than those of the previous experiment, however. We

suspected that one possible reason for this is that in the first experiment, subjects read through all the cards before being asked to sort. In the second experiment, as in Medin et al. (1987), subjects were simply given the cards and asked to sort them into the most natural categories. If subjects did not read over a certain number of the cards, they might be less likely to notice the theme underlying the categories. That is, the sorting task itself might be interfering with subjects' ability to notice the categorical structure, because subjects may start to sort before processing all the stimuli.

To investigate this possibility, we carried out a further study in which we compared the Domain-Consistent and Integrated category structures, crossed with a procedural variable: whether subjects read through the cards before being asked to sort. We expected that previewing the cards would help subjects to identify the category structure in the Integrated condition, but not in the Domain-Consistent condition, since no theme or knowledge structure would come to mind that would distinguish the categories. In fact, this is exactly what we found. In the Domain-Consistent condition, no subject discovered the category structure. In the Integrated condition, the categories were recovered 40% of the time without a preview and 78% of the time with a preview.

This difference between the preview conditions suggests that there is some validity to the analogy drawn earlier of the vacationer who notices the new category of trees. That is, simply looking at the trees may lead one to notice the differences between them, even if one isn't being asked to classify them. When our subjects simply read through the cards, they apparently noticed the structure behind the specific instances. Why are subjects less likely to discover the category when they do not preview the cards? Informal observation suggests that some subjects begin sorting by comparing the first two or three cards. As soon as they discover a difference, they use this to make two piles. This difference can then be used to classify all subsequent cards. Such a strategy is extremely likely to result in a unidimensional sort, because the initial difference will probably be a single dimension, and the subject has not read all the features on a number of cards, which would be necessary to identify the theme. To some degree, then, the bias towards unidimensional sorting which has been extremely robust in the literature (e.g., Ahn & Medin, 1992; Medin et al., 1987) is a function of the sorting task itself. That is, sorting may induce strategies that lead to unidimensional sorts. Looking at the preview condition, we find evidence that category formation can be done without an explicit sorting task, as children appear to do in noticing different types of objects, or as our fictional vacationer did.

This notion led us to perform experiments in which we did not ask subjects to explicitly sort the items at all. Instead, we asked them to carefully read through the cards, and then we asked some increasingly specific questions about what they noticed. Initially, we asked if subjects noticed anything about

the cards (their answers here were seldom helpful—along the lines of "they had something written on them"). Then, we asked whether they had noticed if there were different kinds of things described in the cards. Finally, we asked them to write down the features for the different kinds of things. Once again, we compared the Integrated and Domain-Consistent conditions in this task. Our prediction was that the Integrated subjects might notice the categorical structure simply by virtue of reading through the cards, whereas the Domain-Consistent group, lacking any help from a knowledge structure, would probably not.

There was a difference between the groups when asked whether they noticed two different kinds of things in the cards: 89% of the subjects in the Integrated group noticed a difference, whereas only 33% of those in the Domain-Consistent group did. This dependent measure is a bit suspect, though, because it is very easy for subjects to say they saw two kinds of things. Of greater interest was whether subjects were actually accurate in identifying the features of the categories they claimed to have noticed. We counted the number of subjects who got at least two features correct in each category. (Mentioning only one feature does not demonstrate identification of the category structure; this could have been a unidimensional category.) In fact, 83% of the Integrated subjects provided two or more correct dimensions. However, only 8% of the Domain-Consistent subjects could do so. Thus, when there is a theme unifying the features of a category, people are apt to identify the theme and therefore the category—even when they are not asked to form categories in the task. However, when the features are not related by any particular knowledge, the statistical category structure is apparently not very noticeable.

This is not to say that the structure of the category is unimportant. Obviously, the structure of the stimuli can make the spontaneous formation of categories trivial or impossible. For example, if the two categories consisted of completely different features that never overlapped (e.g., imagine that the two categories were *bananas* and *doctors*), then people would likely notice that they were very different. Similarly, the two categories might have very overlapping structures, such that subjects could only tell the difference by counting up all the features and their co-occurrences. It is very unlikely that subjects would spontaneously notice different categories in such a case.

There are more interesting effects of category structure, as well. In some cases, differences in category structure partly determine whether knowledge will be useful. One illustration of this involves the difference between "cross-over features" and "missing features," both of which are widely used in category learning experiments. Imagine that you are learning about two categories. One category tends to have the features *flies*, *has wings*, *nests in trees*, and *perches on power lines*, while the other category tends to have different values for the same dimensions, *does not fly*, *lacks wings*, *nests in tall grass*, and *perches on the ground*. Now, consider two different structures. In the

first, each instance has three of the four features that are indicative of one of the categories and no information is given with respect to the other feature (i.e., this is the "missing feature"). So, one instance might have *flies, perches on power lines, nests in trees* and another might have *flies, has wings, perches on power lines*, and so on. Here, each instance seems rather consistent with prior knowledge. However, imagine a slightly different structure such that each instance has three features indicating that category and one feature typical of the other category (the "crossover feature"). Now, one of the instances would be *flies, lacks wings, nests in trees, perches on power lines.* This item seems very strange, due to our knowledge of the importance of wings in the ability to fly and the role that flying normally plays in nesting and perching. In some sense, the absence of the wings feature in the first structure is equivalent to the crossover feature in the second structure—both count as evidence against membership in the first category. However, they really do not *seem* equivalent. Intuitively, it seems that it is easier to explain (or maybe ignore) the absence of the feature than the presence of the crossover feature. Changing the structure of the exemplar has changed the instances from being generally consistent with background knowledge to being inconsistent (see also Spalding & Murphy, 1994, for additional discussion of this issue). The point is that, although one cannot discount the content of the features, one cannot be sure whether knowledge can be used to integrate the instances of a category purely on the basis of the knowledge that is related to the individual features. Instead, one must consider both the content of the features and the structure of the category.

Implications for Conceptions of Similarity

Theories of similarity are by their nature highly abstract. If one has a theory of similarity, then it must apply to fish, weddings, beer, tanks, authors, types of soil, etc. The alternative would be to have one theory of the similarity of fish, another for the similarity of weddings, and so on, and this would no longer be a study of similarity itself but a study of fish, weddings, etc. Therefore, theories of similarity have employed very formal principles that can be abstracted across domains, such as the commonality of "features" or the ranking of items along "dimensions" (e.g., Shepard, 1987; Tversky, 1977). That is, these theories do not talk about the meaning of the features or knowledge about the objects being compared. For example, Tversky's (1977) contrast theory of similarity does not refer to particular properties of objects, but only to the number of common and distinctive features of the things being compared. For this reason, it may be applied to any pair of items.

Concept theorists have made an analogous attempt to create general, abstract theories of concepts and categorization, in part based on the assumption

that similarity is the basis for categories. So, theories that claim that people remember exemplars (Brooks, 1987; Medin & Shaffer, 1978) do not specify exactly what is remembered about each exemplar, because this would likely differ with the type of object. Models of prototype structures (e.g., Reed, 1972; Rosch & Mervis, 1975) are also extremely general, allowing one to use the models in a host of different domains. In both cases, the theory can be specified as a kind of representation and learning rule, without reference to the specifics of the concept being learned. The abstractness of these models suggests that the *content* of a given category being learned will have little effect. The work summarized here suggests otherwise. Let us take the category learning case first (Murphy & Allopenna, 1994). In all conditions, the learning stimuli and categorization rule were identical from a formal standpoint. For example, suppose that every feature were replaced by an arbitrary letter of the alphabet (A, B, C...). If we did this for each condition, the learning examples in each condition would be identical: There would be the same number of features, the same number of exemplars, and each learning exemplar in one condition would have an identical exemplar in another condition. Thus, all that differs between the conditions is the content of the features. That is, in one condition, A = "Made in Africa" and in another, A = "Astronauts live there." From the point of view of most theories of similarity, it does not make any difference whether the features are A, B, C... or typographical symbols or English phrases. It is the commonality or differences of the features that counts. Nonetheless, we found considerable differences caused by the content of the features. Furthermore, the differences were predictable based on how consistent the features were with background knowledge.

This result is perhaps even more striking in the case of category formation (Spalding & Murphy, in press). When someone does not know a category, we would expect that the person must use only statistical information to identify it. Using knowledge would seem to assume that the person already knows what the category is, and so would be a circular explanation. Nonetheless, we again found that the content of the features had an enormous effect on subjects' ability to identify the category structure. We discuss this seeming paradox below.

What seems to be going on here is that these knowledge structures may be providing an *explanation* of the features by identifying a theme (whenever possible). If a coherent explanation can be provided for why some features seem to go together but others do not, this may have an important effect on whether the category structure will be identified. Of course, such an explanation can only be arrived at if the features are in some way related to knowledge that the subject already has. The features don't have to perfectly match that knowledge. (In fact, Murphy & Allopenna, 1994, showed that their categories were not familiar to subjects prior to learning.) However, the features do have to be at least inferentially related to the knowledge so that the relations among them can be identified.

In short, a purely formal measurement of conceptual structure would not find any difference between the types of categories compared in these experiments. Nonetheless, these differences had massive effects on category formation and learning. To be fair, formal models of similarity often have a ready response for this problem. They point out that they do not say exactly what the features or dimensions are of every set of stimuli, and so they shouldn't be held to a particular featural analysis. For different stimuli that are formally equivalent (under one analysis), subjects may well perceive different features or dimensions. It is these *perceived* features—not necessarily the features given by the experimenter—that the formal theory uses. Thus, the fact that subjects are not using only the presented stimulus features in categorization makes it difficult to evaluate the model. In short, the attitude is, "If you tell me what the features are, I'll tell you what's similar."

One can use this response to explain why it is that the Integrated Concepts are learned so easily, and why subjects recover their category structure in a formation task. After viewing a few examples, subjects begin to identify the themes. Then they can classify each stimulus as to which theme it is consistent with. Now it is these new features, the themes, that are most heavily weighted in the similarity comparison. Although they are not stimulus features, they are the psychological features that are controlling category formation and learning. With these new features, the formal similarity theory should now be able to explain performance, because such features are assumed not to be present in the Random or Domain-Consistent conditions.

The problem with this position, though, is that all the work seems to be done by the knowledge-based processes that notice and identify the themes. That is, once one has figured out that one kind of object is an underwater building and therefore has thick walls, is inhabited by divers, etc., and the other kind is a building that floats in air and therefore has thin walls, is inhabited by astronauts, etc., the items can be separated by virtue of this simple distinction—one doesn't need fancy rules of similarity or category formation. We are not trying to argue that formal theories don't have any place in explaining how people learn categories, especially in cases where people have little or no prior knowledge about the domain. What we are trying to say is that once knowledge-based processing is allowed into one's explanation of our experiments, there is very little left for the formal models of similarity or categorization to explain. That is, when such theorists say "You tell me the features, and I'll tell you what the category is," they are in a sense asking for the category itself, since subjects are *discovering* or *constructing* the thematic features from the stimulus features while trying to learn the category. Furthermore, one cannot be sure whether a theme can or will be constructed by the subjects simply by considering the meaning of the individual features. As discussed earlier, the construction of a theme seems to depend both on the content of the features and the structure of

the category. Thus, the thematic features should not be presupposed by a similarity theory.

The notion that subjects are constructing the thematic features helps to answer a question about the knowledge-based approach. If knowledge and underlying theories are important to concept learning, then it is natural to ask where they come from. Furthermore, if the knowledge is represented in terms of concepts, then aren't we assuming that the concepts exist before they are learned? Our answer has two parts. First, as we have emphasized already, the knowledge is about a general domain, rather than about the to-be-learned concept. People have knowledge of animals in general, and about some classes of animals, such as herd animals, domestic animals, predators, migrating animals, and so on. So, this knowledge applies across an entire domain or across large classes within it. When learning about a new animal, subjects don't have prior knowledge about that specific kind but do have the domain knowledge that tells them what kind of information to look for and represent. Many of the specific details of the animal must be learned (e.g., its precise shape, coloring and habits), so there is no circularity in the process.

Second, it seems likely that there is considerable interaction between the general knowledge and the specific learning of an individual category. Wisniewski and Medin (1994) suggest that background knowledge suggests a set of hypotheses that are tested and modified in light of evidence from learning. So, if you first think about an underwater house, you might imagine that it is filled with water, and its inhabitants wear scuba gear. However, when you read that it has "thick heavy walls," this can be easily interpreted as a means of keeping out the water, and so now you would conclude that the house has air, rather than water as its atmosphere. Both of these interpretations are consistent with your knowledge of underwater structures (and humans' biological need for air, etc.), but which one is used and the specifics of how it is instantiated depends on the particular features that are actually encountered in the stimuli. Thus, there is an interactive process in which the knowledge suggests interpretations, the evidence reinforces some and rules out others, and the concept is gradually formed by the interaction of the two. As a result of such interaction, learning a new concept can lead to permanent modifications in the knowledge base. The attempt to make sense of empirical data can suggest new knowledge structures that in turn are used to learn new concepts.

In short, we do not make the assumption that background knowledge is a pre-existing structure that rigidly determines the course of concept learning. Instead, there is considerable interaction between the empirical learning aspect and the knowledge-based component. We suspect that much the same process occurs even with young children. Rather than knowledge or formal learning alone being the basis for early concepts, there is probably a process of interaction such that the acquisition of knowledge and the learning of new

concepts develop together. This topic is beyond the limits of the present article, however.

Conclusion

We have claimed that formal similarity, defined in terms of matching and mismatching stimulus features, is insufficient to account for human performance in category learning and formation. The experiments reviewed in this article have shown that the background knowledge brought to bear by the subjects has a huge effect on category learning and formation. Thus, one must consider the category structure, the content of the features and their relationship to knowledge structures in long-term memory in order to understand how people learn or create concepts.

References

Ahn, W. (1990). Effects of background knowledge on family resemblance sorting. *Proceedings of the 12th Annual Conference of the Cognitive Science Society* (pp. 149-156). Hillsdale, NJ: Erlbaum.

Ahn, W. (1991). Effects of background knowledge on family resemblance sorting and missing features. *Proceedings of the 13th Annual Conference of the Cognitive Science Society* (pp. 203-208). Hillsdale, NJ: Erlbaum.

Ahn, W., & Medin, D. L. (1992). A two-stage model of category construction. *Cognitive Science, 16*, 81-121.

Brooks, L. (1987). Decentralized control of categorization: The role of prior processing episodes. In U. Neisser (Ed.), *Cognition and conceptual development: Ecological and intellectual factors in categorization* (pp. 141-174). New York: Cambridge University Press.

Clark, E. (1981). Lexical innovations: How children learn to create new words. *Behavioral Development: A Series of Monographs*. 299-328.

Homa, D., Rhoads, D., & Chambliss, D. (1979). The evolution of conceptual structure. *Journal of Experimental Psychology: Human Learning and Memory, 5*, 11-23.

Hull, C. L. (1920). Quantitative aspects of the evolution of concepts. *Psychological Monographs, 28*.

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist, 44*, 1469-1481.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100*, 254-278.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987) Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology, 19*, 242-279.

Merriman, W. E., Schuster, J. M., & Hager, L. (1991). Are names ever mapped onto preexisting categories? *Journal of Experimental Psychology: General, 120*, 288-300.

Mervis, C. B. (1987). Child-basic object categories and early lexical development. In

U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 201-233). New York: Cambridge University Press.

Murphy, G. L. (1993a). A rational theory of concepts. In G. V. Nakamura, R. Taraban, & D. L. Medin (Eds.), *The psychology of learning and motivation: Vol. 29. Categorization by humans and machines* (pp. 327-359). San Diego: Academic Press.

Murphy, G. L. (1993b). Theories and concept formation. In I. Van Mechelen, J. Hampton, R. Michalski, & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis* (pp. 173-200). London: Academic Press.

Murphy, G. L., & Allopenna, P. D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 904-919.

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

Pazzani, M. J. (1991). Influence of prior knowledge on concept acquisition: Experimental and computational results. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 416-432.

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

Rosch, E., & Mervis, C. B. (1975). Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573-605.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological space. *Science, 237*, 1317-1323.

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

Spalding, T. L., & Murphy, G. L. (in press). Effects of background knowledge on category construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

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

Wattenmaker, W. D., Dewey, G. I., Murphy, T. D., & Medin, D. L. (1986). Linear separability and concept learning: Context, relational properties, and concept naturalness. *Cognitive Psychology, 18*, 158-194.

Wisniewski, E. (1995). Prior knowledge and functionally relevant features in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 449-468.

Wisniewski, E., & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science, 18*, 221-281.

Received October, 1994