Knowledge Effects

Traditionally, studies of concepts have used materials that are as divorced as possible from outside knowledge. A survey of the literature would find stimuli such as geometric shapes, alphanumeric strings, patches of color, dot patterns, and schematic faces to be quite common. Although some of these stimuli have familiar elements, such as faces or geometric shapes, people's knowledge prior to the experiment could not have predicted what the concepts were. For example, nothing that one knows about geometry suggests that the to-be-learned concept should be red pairs of triangles, rather than single blue circles. Why are such categories so popular? There are two main reasons. First, researchers believe it is important in a learning experiment to use categories that the subjects do not already know. If subjects are already familiar with the materials, then the results would not just reflect learning but also their pre-experimental knowledge. In a study of category learning, then, artificial stimuli may be best. Second, by using very simple, even meaningless stimuli, some investigators feel that they are discovering general principles that will apply across many different domains. If the stimuli had been types of animals, for example, how could we be sure that the results would generalize to plants or furniture or economic principles? But when the stimuli are abstract items that are not related to any particular domain, the results may reflect basic principles of learning that would apply very widely.

It is not obvious that this reasoning is actually correct, however. It could be, for example, that the results of such experiments do not apply to more realistic situations in which people know something about the domain and do have expectations, right or wrong, about the concept. And just because the stimuli are simple and abstract does not mean that the results are especially generalizable. The history of cognitive psychology is based, one could argue, upon the finding that results from simple learning experiments (on rats, pigeons, or aplysia) do not give much insight into intelligent behavior and thought. Whether experiments on simple
categorization could be the basis for explaining all concepts is an empirical question. However, if one is worried that experiments on learning animal categories would not generalize to learning about furniture, then how can we be sure that experiments on learning dot patterns generalize to both of them?

In contrast to the traditional assumptions, one could argue that the learning processes used for realistic concepts in familiar domains are not the same as those used for simple, artificial situations. It is not (or should not be) controversial to point out that there are important variables in real-world learning that are never identified or explored in the simple experiment, because they have been “controlled away” by the experimental design. In a more controversial sense, the results of the simple experiments may be actually wrong, in that people could be doing things in those situations that they do not do in real life. It is important to emphasize, however, that these are empirical questions. Some people love the elegance and control of the simple experiments, and some people find them boring. Discussions of which approach is correct have an unfortunate tendency to degenerate into personality clashes. This issue must be resolved not by argument but by research into what people do in the richer situations, which can then be compared to the results of the majority of past experiments on concepts, which follow the simple, abstract procedures. It is possible, and perhaps likely, that both approaches to understanding concepts will be necessary to paint the complete picture.

The present chapter, then, provides a review and discussion of a class of effects found in more realistic stimuli, in which some (but by no means all) of the richness of real-world stimuli is allowed into the laboratory and is examined. Specifically, this chapter examines the effects of background knowledge on concept learning and use, an issue that has not been a traditional part of the psychology of concepts but which has slowly grown in popularity since the mid-1980s (Carey 1985; Keil 1989; Murphy and Medin 1985).

Possible Benefits and Limitations of Prior Knowledge

Previous chapters have already reported a number of benefits of prior knowledge in concept learning and use. Nonetheless, it would be useful to step back and discuss some general ideas of why and how knowledge might be useful—or not—to put the whole enterprise into perspective. To start very simply, how might knowledge be related to or useful in concepts? One way some people summarize this possibility is to say something like “People have a theory of dogs, which is their concept of them.” This gives the impression that the knowledge is sufficient to predict the existence of dogs and/or to account for all their properties, just as a scientific theory
explains the existence of black holes, say. Such a situation is not impossible, as in some sciences it is possible to predict entities before they are actually seen. For example, Neptune was predicted to exist before it was actually discovered; and in subatomic physics, various particles have been predicted long before experiments could be devised to actually create and then detect them. However, I doubt very much whether people's knowledge is this powerful in most cases. If you have mundane knowledge of the seas, for example, it would not form a theory that would allow you to predict the existence of dolphins or sea slugs or sponges if you didn't already know that they existed. Instead, knowledge is almost always used in a post-hoc way to explain entities and apparent categories that one encounters. For example, if you saw a sea slug at the bottom of the ocean, you could use your knowledge to understand what kind of thing it was and what properties it probably has (e.g., gills rather than lungs). Therefore, I will focus here on the use of knowledge to explain a category when it is being learned, and use that as a framework to understand knowledge effects. (This discussion draws on material in Murphy 2000.)

Let's consider how a common category, birds, might be explained to some degree by everyday knowledge. Most people think of birds as being feathered, two-legged creatures with wings, which fly, lay eggs in nests, and live in trees. Why do birds have these properties? In asking this, I am not asking a question of evolution but of understanding why this particular configuration of properties exists rather than some other. With simple, mundane knowledge, one can explain many of these features. Let's start with flying. In order to fly, the bird needs to support its weight on wings. The feathers are important as a very lightweight body covering that also helps to create an aerodynamic form. Thus, wings and feathers enable flying. By virtue of flying, the bird can live in nests that are in trees, because it can easily fly into and out of the trees. This is a useful thing to do, because many predators are unable to reach the nests there. The bird needs a nest for brooding, and for the babies to live in until they are able to fly. Thus, flying can be partly explained by these desirable consequences.

This line of reasoning, which virtually any adult in our culture could perform, relies not on book-learning or courses but on everyday knowledge. This knowledge may be incomplete or even wrong in detail (e.g., most people's ideas about how wings support a bird are probably wrong), but it is close enough to account for a number of generalizations about the world. For example, most things that fly have wings; almost everything with wings flies. The exceptions can be explained by the mundane knowledge as well. For example, ostriches have wings and do not fly, but this is easily explained by the tiny size of the wings relative to the size of the ostrich.
The surface area of the wings is manifestly unable to lift the ostrich's weight. So, even if people do not really understand the aerodynamic principles by which wings support a body, what they do understand (or believe) about them does a pretty good job in explaining why some things can fly and some cannot.

All I am claiming is that these particular properties of birds can be explained once one has noticed them. It might be possible for a physiologist, for example, to know some of the features and then to be able to predict others (e.g., predicting the internal structure of birds' bones, given observations of their flying and wing structure), but such a feat would be far beyond most of us. Instead, most people can provide a post-hoc explanation of why birds have the particular constellation of features they do. The post-hoc nature of such explanations makes them somewhat circular. One cannot just explain why a bird has wings. One must explain why it has wings, given that it flies, lives in a nest, lays eggs, and so on. But at the same time, the explanation for flying depends on knowing that the bird has wings, lives in nests, lays eggs, and so on. Rather than a logical chain in which unknown properties are deduced from known ones, each property can be explained by the other properties. These properties are in a mutually reinforcing or homeostatic relationship such that they conspire to support one another (Boyd 1999; Keil 1989). Such circularity would be anathema in science, but from the perspective of learners, circular explanations may still be useful.

There are some exceptions to the post-hoc nature of explanations. For example, if someone told you about a DVD player, they might say “It's like a CD player, but for movies.” Or perhaps when you first saw a DVD inserted into a player, you thought something like that. Now you can bootstrap a concept of DVD players by drawing properties that you know from this related concept. However, this is only successful if you are told or observe the relationship between DVD and CD players—it is rather specific similarity between two categories rather than more general domain knowledge. Although this kind of similarity is probably very useful in many cases, I will be focusing in this chapter on broader domain knowledge.

Another shortcoming of explanations is their shallowness (Wilson and Keil 2000). Often one can explain a given property by referring to some underlying property or principle. But why that principle exists or where that underlying property came from is often totally opaque. For example, I can explain the fact that knives are made of metal by the fact that metal is hard and therefore helps the knife in cutting. But why is metal hard? Although I know that it is somehow based on chemical-physical properties, I don’t really have a clue what they are. (NB: It is not necessary
to write and explain them to me.) This shallowness of explanations seems typical. Like the doctor in Molière's *Le Malade Imaginaire* who explained the effect of sleeping pills as arising from their "dormative virtue," we don't understand many properties more than one level down.

The question, then, is whether such knowledge will in fact be useful. If explanations are often applied after the fact and are rather shallow, can they really aid in learning and using concepts? Specifically, if one typically explains features after observing them, does knowledge arrive too late to help in initial learning? If one has to know that a bird has wings and flies in order to understand the connection between them, it looks as if one has already learned the critical features of birds. The explanation might make one feel a sense of understanding after the fact, but it may not have influenced the actual learning of the properties. If explanations are generally somewhat shallow, then how helpful could they be in using a concept, in tasks such as induction or communication?

In short, although it may seem intuitively obvious that our knowledge of the world will help us learn new concepts related to that knowledge, it is possible that the limitations on our knowledge will make it less than useful in many cases. Among those cases would be situations in which knowledge is incomplete or partially wrong, unfortunately not unusual circumstances for humans.

**An Example of Knowledge Use**

Before discussing actual experiments, let's consider a somewhat simplified example of how you might normally learn a new category, namely a trip to the zoo. This situation can be used to illustrate the specific uses of knowledge to be discussed and can be contrasted with the typical psychology experiment using artificial stimuli. Suppose that on a trip to the zoo, you encounter a new animal, the whingelow. The whingelow, let us hope, does not look anything like a dot pattern, a geometric figure on a card, or a string of letters. Instead, the whingelow probably looks much like other animals of its class. If it is a mammal, it likely has visible fur, four limbs, a symmetrical face with two eyes, a nose, two ears, and so on. Thus, the whingelow is unlike the stimuli in these experiments in just the respect that it *does* look like other kinds of things you know. Furthermore, if you find out (or can infer) that it's a mammal, then you can use your knowledge about other mammals to conclude that it has a four-chambered heart, gives birth to live young, and so on. This real-life situation, then, is exactly what the experimentalists described above were trying to avoid: The real-life concept learner is not in a pristine environment where past
knowledge is useless—instead, the learner comes to the situation knowing much about similar kinds of animals and being able to use that knowledge to learn about the whinglew.

It must be emphasized that before your trip to the zoo, you did not already know what the whinglew was, perhaps never having heard about it. You might have known a lot about animals and mammals in general, however. So, this prior knowledge is knowledge about the entire domain (e.g., all animals must breathe) or about concepts related to the new one (e.g., the whinglew looks a bit like a wombat; animals in zoos are often endangered), rather than knowledge about the new concept itself. The question, then, is whether and how such knowledge might influence your learning about whinglew.

Before beginning, however, it is necessary to clarify some terminology. I refer here to **knowledge effects** as influences of prior knowledge of real objects and events that people bring to the category-learning situation. The word “knowledge” is polysemous, and it could also be used more broadly than I am here to indicate anything at all that one knows about a category. For example, once a subject in Posner and Keele’s (1970) experiment had learned the dot-pattern categories, she could be said to have “knowledge” of the categories. But this is not the sort of knowledge that we are discussing in this chapter. The knowledge that we will investigate is the pre-experimental information about the world, rather than what you know about the target concept in particular. (Sometimes I shall use the phrase background knowledge, which emphasizes this point.) As we shall see, such knowledge can have a powerful effect on learning.

Knowledge effects are to be contrasted with **empirical learning** of the properties of observed exemplars during learning. For example, at the zoo, you might see that most whinglewos have an eye stripe; or you might remember a purple, mid-sized whinglew with two legs that whined. These observations would constitute empirical information about the category.

Researchers have suggested that knowledge is important in a number of different aspects of category use (Heit 1997; Murphy 1993; Murphy and Medin 1985). First, knowledge may be involved in defining the features of an object. For example, when you went to the zoo, you paid attention to the number of legs of the whinglew and incorporated this into your initial concept. But you didn’t incorporate the location of the animal’s cage or the time of day. These didn’t seem to be important aspects of the concept of a whinglew, even though they were readily perceivable properties. It can be argued that it was your prior knowledge of animals that directed these judgments about what features to encode and use. In other domains, the features you
ignored for whingelows might be very reasonable ones; for example, social events may exist at specific places or times of day even though animal concepts generally do not. Thus, one’s domain knowledge could cause one to encode or ignore time of day when learning the new animal concept.

Second, knowledge could help people to learn the features of new categories. It is well known from memory experiments that it is difficult to learn arbitrary lists of things, but once those lists are formed into a coherent structure (e.g., a story or description), people learn the material much better. Thus, it could be easier for you to learn about the whingelow than about Rosch and Mervis’s (1975) alphanumeric strings, because your prior knowledge about animals may aid learning the particular features of the whingelow.

Third, your knowledge could influence the kinds of categorization decisions you make after learning. For example, if you saw a smaller animal that was in the same cage as the whingelows, and that was hanging around one of the females, you might assume that this was a baby whingelow, even if it did not look identical to the adults (Mervis 1987). This is not due to your having learned (yet) what the babies look like, but instead would be explained by your broader beliefs about the behavior of juvenile animals. More generally, the way you categorize new objects may be partly dependent on your background knowledge, rather than solely on your experience with the category.

Finally, knowledge might also be used to guide inductions or inferences you make about a category. For example, if one of these (apparent) baby whingelows kept pestering one of the females, you might infer that it was attempting to suckle. Again, such an inference is not based on specific experience with that animal (or even with whingelows) but is an educated guess, based on your more general knowledge. This chapter will discuss each of these uses of knowledge in turn.

Knowledge Effects in Category Acquisition

Category Learning Experiments
Perhaps the simplest way to start this investigation is to ask, “Does it make any difference what you know when you are learning a new category?” Intuitively, the answer seems obviously to be “yes,” but this answer is not one that you would receive from most theories of concept learning. For example, suppose that concept learning consists in forming a prototype, which is the set of features that are most distinctive of the category. Nothing in this process relies on knowledge—it is a matter of identifying the features of the category members and seeing which ones
are the best indicators of the category (see chapter 3). This is a statistical process that does not depend on the identity of the features. Similarly, if learning the category amounts to learning individual exemplars, there is no obvious reason that background knowledge should be involved. And in fact, the main articles describing prototype and exemplar theories say little if anything about how background knowledge might be involved in category learning (Hampton 1979; Medin and Schaffer 1978; Nosofsky 1984; Reed 1972; Rosch 1973; Rosch and Mervis 1975). This is not to say that these authors would actively deny that knowledge is important but to point out that although we may have an intuitive belief that knowledge should influence category learning, the most prominent theories of category learning have not incorporated this idea.

A simple demonstration of the importance of background knowledge can be found in Murphy and Wisniewski's (1989) Experiment 4. In one condition, subjects learned Coherent categories. These were items in which the features were sensibly related to one another, or at least not contradictory. For example, subjects learned an animal category that usually had the features: lives in water, eats fish, has many offspring, and is small. In contrast, the Incoherent categories had features that did not make sense together, for instance, lives in water, eats wheat, has a flat end, and used for stabbing bugs. (These features were used to construct individual items and were presented as verbal descriptions. The details are not essential here, except to note that the concepts were formally identical—only the names of the features changed, not the structure of the exemplars or the feature distributions.) As can be seen, in the Incoherent condition, the features came from different kinds of domains (animals and tools), and even within those entities, they did not make much sense (if something is used to stab bugs, it should have a pointy end, rather than a flat one). And perhaps not surprisingly, Murphy and Wisniewski found that the Incoherent categories were learned less well than the Coherent ones. Subjects made more errors on test items, and they were less confident in their categorization decisions for Incoherent categories.

In short, a concept's content influences how easy it is to learn. If the concept is grossly incompatible with what people know prior to the experiment, it will be difficult to acquire. This conclusion, however, is not very impressive, since most concepts that people learn are not nearly as incoherent as those in the Incoherent condition. Pazzani (1991) asked a more interesting question, namely whether prior expectations could help people learn a category compared to a case in which they had no such expectations. Here neither category was incompatible with what is already known; one category was consistent with prior knowledge, and one was simply not related to any particular knowledge.
Pazzani (1991) used categories that were described by simple rules. His items were pictures of people performing actions on objects. Each picture showed an adult or child doing an action on an uninflated balloon that was either large or small and either yellow or purple. The action was to dip the balloon in a cup of water or else to stretch it. Pazzani compared two different kinds of categories, disjunctive and conjunctive. The disjunctive category was defined by the rule “the person must be an adult OR the action must be stretching the balloon.” The conjunctive category was defined by the rule “the color must be yellow AND the balloon must be small.” Considerable past research had shown that people find it easier to learn a conjunctive category than a disjunctive category, so that is what one would expect in this case as well. Pazzani added another factor, however, which he believed would influence the result. He told subjects either that they were to try to learn “Category Alpha” or that they were to try to identify which balloons would inflate. He believed that the first condition would not activate any particular knowledge but that the second condition would bring to mind whatever people believe about inflating balloons. In particular, people may believe that adults are better able to inflate a balloon than children are and that stretching a balloon makes it easier to inflate. Thus, this knowledge would be expected to aid the disjunctive rule, which included these features. Subjects saw the pictures one at a time and had to say whether each one was in the category that they were learning. The dependent measure in the experiment was how many examples of the category they had to see before they were completely accurate in identifying category members.

When subjects received the Category Alpha instructions, Pazzani found the usual pattern of the conjunctive category being easier to learn than the disjunctive category (a difference of about 15 fewer trials). However, when the Inflate instructions were given, this result reversed; in fact, the disjunctive category was learned after only about 8 examples on average, whereas the conjunctive category was learned after about 28 examples (Pazzani 1991, figure 1). Clearly, the hint about balloon inflation greatly aided subjects in learning the category. Knowledge was useful in a more positive way than was demonstrated by the Murphy and Wisniewski experiment—namely, people used their knowledge to make hypotheses about a category’s features, and since their knowledge was correct, concept learning was easier.

Pazzani’s (1991) experiment has another implication as well. It is significant that a finding from many past experiments on category learning (starting with Bruner, Goodnow and Austin 1956), the advantage of conjunctive rules, could actually be reversed by the use of background knowledge. This suggests the somewhat disturbing possibility that the results of artificial experiments that do not make contact with prior knowledge may not hold in the situation where people have expectations
about the category (which is probably the more usual case). That is, it is possible that results from experiments with dot patterns or geometric shapes will not apply to real-world cases in which people have background knowledge. This possibility will be discussed further in the chapter’s conclusion.

Locus of the Knowledge Effect
Clearly, then, knowledge can help one learn a novel category. But there is considerably more to be investigated here. If one compares the traditional experiments with artificial stimuli (such as dot patterns) to natural categories, two differences are salient. The first is that the features themselves are different. The second difference is that the features of natural categories seem to hang together, whereas the features of artificial categories generally do not. In order to understand the knowledge effect, we need to discover whether either or both of these differences is responsible for it. The following experiments investigate these variables.

One possible reason for knowledge effects is that the features themselves may make learning easier. As noted earlier, in most experiments that do not involve knowledge, the stimuli have been themselves meaningless items, such as dots, geometric stimuli, and even letter strings (individual letters not having much meaning associated with them). Would using more meaningful stimuli in and of itself lead to an improvement in learning? Murphy and Allopenna (1994) compared categories that had relatively meaningless typographical symbols (like $<$, $\{$, $+$, and $!$) as features to those that had meaningful phrases such as “lives alone” and “thick, heavy walls.” However, these phrases were randomly assigned to categories, so that there was no overall rhyme or reason to them. Murphy and Allopenna did not find any difference in the difficulty of learning these categories; both were quite difficult. Thus, the mere meaningfulness of the features itself does not appear to be a very important aspect of the knowledge effect.

What did make an important difference in learning was how the features were related. When the features of a category formed a consistent set, the category was much easier to learn than when they were inconsistent or simply neutral. Consider the pair of categories summarized at the top of table 6.1. The features of these categories, both vehicles, can be described by a theme that connects them. The category in the left column seems to be a kind of jungle vehicle, and the one in the right column seems to be an arctic vehicle. That information was never given to subjects, but it is possible that during the course of learning they would realize that the features were related in this way, and this might in turn improve the learning process. Because these features could be integrated into a single theme, Murphy and Allopenna
Table 6.1.  
Typical features of categories from Murphy and Allopenna (1984).

<table>
<thead>
<tr>
<th>Integrated Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category 1</td>
</tr>
<tr>
<td>Made in Africa</td>
</tr>
<tr>
<td>Lightly insulated</td>
</tr>
<tr>
<td>Green</td>
</tr>
<tr>
<td>Drives in jungles</td>
</tr>
<tr>
<td>Has wheels</td>
</tr>
<tr>
<td>Category 2</td>
</tr>
<tr>
<td>Made in Norway</td>
</tr>
<tr>
<td>Heavily insulated</td>
</tr>
<tr>
<td>White</td>
</tr>
<tr>
<td>Drives on glaciers</td>
</tr>
<tr>
<td>Has treads</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Neutral Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category 1</td>
</tr>
<tr>
<td>Green</td>
</tr>
<tr>
<td>Manual transmission</td>
</tr>
<tr>
<td>Radial tires</td>
</tr>
<tr>
<td>Air bags</td>
</tr>
<tr>
<td>Vinyl seats</td>
</tr>
<tr>
<td>Category 2</td>
</tr>
<tr>
<td>White</td>
</tr>
<tr>
<td>Automatic transmission</td>
</tr>
<tr>
<td>Non-radial tires</td>
</tr>
<tr>
<td>Automatic seat belts</td>
</tr>
<tr>
<td>Cloth seats</td>
</tr>
</tbody>
</table>

Note: Subjects would have learned a pair of integrated or neutral categories. Exemplars were constructed by selecting a subset of each category's features (shown above), along with some random features (not shown) that occurred in both categories equally often.

called this the Integrated Condition. Now consider the pair of categories summarized at the bottom of table 6.1. These categories are also vehicles, but their properties are no longer connected by a theme. There is nothing inconsistent about the features that appear together in the category—but there is no special connection between them, either. That is, there is no reason why being green should lead a vehicle to have a manual transmission rather than an automatic one, nor why air bags should go with vinyl seat covers rather than cloth seat covers. I will call this the Neutral Condition. The category structures made from these features were identical—all that differed was how well the features could be related by a theme. The results showed that subjects could learn the Integrated categories in about half the time of the Neutral categories.

These findings suggest that knowledge helps learning because it relates the features in the category, rather than through the properties of the features themselves. Of course, it is likely that some differences in features are important, but more significant here is that the knowledge must relate the features to one another, and hence, to the category as a whole. Why is this helpful? Keep in mind that not every
example from the category has the same features—one might have treads and be white, while another might be built in Norway and drive on glaciers. This variation in features makes the purely empirical learning process more difficult. One must learn multiple features, because every feature will not be present in every item, and an item might have an atypical feature. Learning these multiple features is easier if they can be related by a common knowledge structure. The knowledge could help subjects learn which features go with which categories and to remember that feature assignment at test. If you know that vehicle X is built in Norway, it’s easy to learn that vehicle X goes on glaciers too.

In short, knowledge probably works by helping to relate features. Later, we will consider how using this knowledge affects the category representation.

Amount and Consistency of Knowledge

In the experiments discussed so far, the knowledge has generally been consistent with the entire category. For example, in Pazzani’s (1991) experiment, the category was defined by the two features “adult OR stretch the balloon,” and it was exactly these two features that were related to people’s prior knowledge about inflating balloons. And in the Murphy and Alloppena experiments just described, all of the category’s features were related to the theme. But this situation seems rather unrealistic. When you went to the zoo and saw the whingelow, you had some knowledge that explained some of its properties or that helped them make sense to you. But certainly others of its properties were not specifically related to that knowledge. The particular color of the whingelow, for example, or the shape of its nose, or the length of its fur might not have been related to any specific knowledge that you had. Yet, you might well learn from this experience that the whingelow is a kind of grayish animal, with a wide nose and short fur. Thus, even though you had knowledge about other animals and mammals that was relevant to learning about the whingelow, you probably also learned about properties that did not make contact with that knowledge. In the experiments discussed so far, there were no such properties. This raises the question, then, about whether knowledge is helpful in the more realistic situation. Perhaps when knowledge does not pick out most or many of the category’s properties, it is not very useful.

Kaplan and Murphy (2000) investigated the situation in which some, but not all, of a category’s features were not related to knowledge. Their categories were for the most part like the Neutral categories shown in table 6.1. However, every item had exactly one property that was related to a theme, such as arctic vehicle or jungle vehicle. These properties were different across items: One item had “green,” another
had “drives in jungles,” a third had “is made in Africa,” and so on. Thus, unlike the
Pazzani or Murphy and Alloppenna experiments, the majority of the features of every
item were not related to specific knowledge. Furthermore, in order to realize that
there was a theme related to the categories, subjects had to notice the relatedness of
different features appearing in different exemplars (e.g., that one item’s “green” was
related to another item’s “drives in jungles”). In previous experiments, the exact
same feature was repeated across items (e.g., “stretch the balloon” in Pazzani’s ex-
periment). In spite of these apparent hindrances to noticing and using the knowl-
edge, Kaplan and Murphy’s subjects learned the categories significantly faster (in
less than half the time) when these thematic features were present. Thus, we can
conclude that background knowledge is helpful even when it is not related to all—
or most—of a category’s properties. This is a significant result, because it suggests
that knowledge is likely to be useful even in complex settings in which one’s
knowledge is incomplete or imperfect.

Similarly, Murphy and Kaplan (2000) found that knowledge was useful even when
it was not totally consistent. One popular way of making up items in a category-
learning experiment (Medin, Wattenmaker, and Hampson 1987) is for each item to
consist of N features, of which 1 feature is “incorrect” and N − 1 are correct. (As
shown in table 5.2 of ch. 5, except for the first exemplar in each category.) For ex-
ample, if there are six stimulus dimensions, each exemplar would have five correct
features and one incorrect feature. By “correct” and “incorrect” here I mean simply
that the feature is normally associated with the exemplar’s category or a different
category, respectively. To use a real-life analogy, I might make up an item that has
wings, walks, lives in nests, lays eggs, and sings. Five of these features are associated
to the category of birds, and one is associated to mammals. If the features in an ex-
periment are derived from a theme, like those shown in table 6.1, how does the
presence of the incorrect feature in each item affect learning? Because every item has
a feature that is inconsistent with the rest of them, perhaps subjects will ignore the
theme; perhaps the knowledge has to be almost perfectly consistent before subjects
will rely on it.

Murphy and Kaplan found that even with this structure, Integrated Categories
were easier to learn than Neutral Categories. Subjects were able to explain away or
ignore the one inconsistent feature and still use the mostly accurate knowledge.
Thus, knowledge does not have to be related to every feature in the category, and it
does not have to be perfectly reliable in order to benefit learning.

Finally, it may be useful to ask how soon knowledge has its effect. Heit and
Bott (2000) propose that knowledge will only be useful after a certain amount of
learning, because it will take a fair amount of time to figure out which knowledge is relevant and is reliably associated to the categories being learned. And in their first experiment, they found no difference between knowledge-related and -unrelated features after one block of learning, but this difference did appear in later blocks. However, Heit and Bott's second experiment with different materials did find a difference after one block. Furthermore, Kaplan and Murphy (2000) also found a difference after one block of learning in their categories with minimal knowledge. In that case, each knowledge-related feature had been seen only once, and yet they were learned better than the more frequent knowledge-unrelated features. It is likely that the time course of knowledge effects will depend greatly on the particular kind of knowledge present in the stimuli: A hint given in advance could affect the processing of the very first stimulus; a subtle knowledge difference that is not readily apparent could take a few blocks to reveal itself. However, there is certainly no general rule that considerable learning is required before knowledge begins to help.

So far, knowledge has been shown to aid in learning categories. However, background knowledge cannot be expected to always be correct. In some cases it may be vague, in others, wrong. (I am flouting convention here by referring to incorrect knowledge, which is an oxymoron. The point is that this information is exactly the same as information that is helpful—only it is wrong.) Clearly, knowledge may not be helpful in such cases. The degree to which "bad" knowledge hurts has not been very much investigated. Kaplan and Murphy (2000) found that when knowledge in a category was contradictory (i.e., features related to different themes were mixed together), subjects found the category about as easy to learn as categories in the Neutral condition (see bottom of table 6.1) that had no knowledge at all. Although the wrong knowledge obviously did not improve learning, it didn't hurt it relative to no knowledge (see also Murphy and Alloppena 1994). This rather surprising result has not yet been fully investigated. The best guess is that when the knowledge is contradictory or misleading, subjects very quickly realize this and then simply ignore it. Thus, even though knowledge is generally helpful, people can apparently identify when it is not, by noticing internal inconsistencies and by receiving feedback.

In a series of experiments, Heit (1994, 1998) investigated how prior knowledge influences the way that new category examples—consistent or inconsistent with one's expectations—are processed. He posed the following thought experiment. Imagine that you are visiting a new country and have read some guide books on the people and sights there. The guide book has given some presumably stereotypical information on the character of the people of this country—say, that the people are typically
friendly. When you go to the country, you will meet a number of people, and some of them are likely to be friendly and others not so friendly. How do you evaluate this information in coming to a conclusion about the character of the people? Given that you expected the people to be friendly, perhaps you should give greater weight to the people who meet this expectation—after all, the unfriendly ones are likely to be exceptions. Alternatively, perhaps you should give greater weight to the disconfirming information. These unfriendly people might surprise you and therefore draw attention to themselves. Thus, they might have a greater effect on your judgment about the people.

Heit’s experiments followed this example in presenting to subjects descriptions of people in a new city, W. He manipulated subjects’ expectations by using pairs of characteristics that would be expected to co-occur or not. For example, if a person was described as “shy,” then one would expect the person to have the property “does not attend parties often” rather than “attends parties often.” Heit manipulated how often pairs of features like “shy” and “does not attend parties often” co-occurred: They could be paired 0%, 25%, 50%, 75%, or 100% of the time. After viewing a number of these descriptions, subjects were asked to estimate percentages of the co-occurrences. For example, they were asked “if a person from city W was shy, how likely would the person be to attend parties often?” based on the exemplars.

Heit expected that subjects would be influenced by their prior knowledge in making these judgments. That is, they should give higher percentages for shy—does not attend parties often than for shy—attends parties often. Figure 6.1 illustrates this effect, by showing that high expectancies (q = 90%) lead to higher estimates than low expectancies (q = 10%) across the range of observed probabilities (the x-axis). However, the exact form of this effect should vary depending on which kind of example subjects pay more attention to. If people pay equal attention to examples that are consistent and inconsistent to their knowledge, then the effect should be equal across the observed co-occurrence of the pairs of features (as in figure 6.1a). If the congruent items (those consistent with knowledge) are weighted more, then the difference between consistent and inconsistent items should be largest at 50% and smallest at 0% and 100% (figure 6.1b). Because the evidence is most mixed at 50%, there is more “room” for knowledge to push up the congruent items and push down the incongruent ones. In contrast, if people attend to the incongruent items (those inconsistent with prior expectations), then the effect should be smallest at 50%. Attention to incongruent items would reduce the difference between the two item types, by pushing up their subjective frequency. The effect would again be found in
Figure 6.1
Predictions from Heit's (1994, 1998) weighting model. Each graph shows subjects' estimated probabilities of co-occurrence of pairs of items as a function of their actual co-occurrence (shown on the x axis), knowledge status (pairs expected to co-occur are solid lines; pairs expected not to co-occur are dashed), and form of weighting. In the equal weighting model, positive and negative exemplars are treated as equally informative. In the congruent model, positive evidence is weighted more heavily than negative evidence; in the incongruent model, negative evidence is weight more heavily. The congruent model shows the largest expectation effect when observed proportions are intermediate, whereas the incongruent model shows the smallest effect at intermediate proportions.
the middle, rather than near 0 or 100% (figure 6.1c), because it makes little difference how items are weighted when almost all the items are of one type. (Heit 1998 provides a mathematical model of the weighting process to support these predictions, which the interested reader should consult.)

Across many experiments, Heit found that congruent and incongruent items were weighted equally—that is, the effect of knowledge was constant from 0–100% co-occurrence (like figure 6.1a). However, Heit (1998) found one exception to this rule: When subjects were given a small number of items and a fairly long time to study each one (16 seconds), they tended to weight the incongruent items more, as shown in figure 6.1c. Here it can be seen that there is an overall knowledge effect; people tend to give lower ratings to the less expected feature pairs. However, this effect is smallest at the middle range of probabilities, because the incongruent items are weighted more (raising that curve) and the congruent ones less (lowering its curve).

Heit (1998) argued from these results that when people encounter an unexpected item, they spend more time thinking about it, perhaps attempting to explain its peculiar properties. And, indeed, he found that people studied the incongruent items longer when the study period was self-paced. This, then, is another effect of prior knowledge. It does not just aid category learning when it is consistent with the examples (as in the above experiments), but it also influences what subjects attend to when learning. It may seem paradoxical that it results in attention to items that disconfirm the knowledge, but this is probably a more efficient strategy for learning. There is no need to give special attention or weight to things that one already knows. Indeed, it is probably best not to think about them, since they do not teach anything new (see also Kruschke 1991). Instead, learning may be most efficient when it focuses on surprising events.

This section began with the question about whether knowledge would be helpful when it was not as complete and reliable as in the early experiments that investigated knowledge effects. The clear answer is “yes.” When knowledge relates only some of the features, it still benefits learning; and when the knowledge was sometimes wrong, it still helped. In Heit’s experiments, subjects were sensitive to both their prior expectancies and the observed level of co-occurrence of features—incorrect knowledge did not overwhelm the empirical data, nor was it ignored. Furthermore, even when the knowledge was plain wrong, it did not adversely affect learning, compared to no knowledge at all. Thus, the picture given by these studies is one in which learners use whatever amount of knowledge is present and useful but are not overly swayed by inaccurate expectations.
Interactions of Structure and Knowledge

In some cases, one might think of knowledge effects as being something that can be added to or subtracted from a category, much as one could print the stimuli in blue or red. And, indeed, many studies have added or subtracted knowledge by changing some features or prior information about the category. However, in an important study, Wattenmaker et al. (1986) argued that knowledge must be related to the structure of a category in order to be helpful. To see what they meant by this, I discuss their experiments in some detail.

Wattenmaker et al. began with the comparison of linearly separable and nonlinearly separable categories described in chapter 4. To summarize this distinction briefly, linearly separable categories are those in which the evidence for the category can be simply summed together in order to make a decision. For example, if the members of category A are mostly blue, triangles, and tall, then one can assume that a blue, tall rectangle would be a member of category A, because the majority of its features are consistent with the category. (This is simplifying somewhat—see chapter 4.) In contrast, a nonlinearly separable category cannot be explained by such a rule. There is no way to simply sum up the evidence for individual features to decide whether the item is in the category. Typically, such categories are determined at least in part by specific configurations of features (e.g., blue-and-triangle might be associated to the category, although blue alone is not).

Table 6.2 gives an example of linearly separable and nonlinearly separable categories (from Wattenmaker et al. 1986). (Each row describes a category exemplar; the columns refer to different stimulus dimensions; and each 1 or 0 indicates a different value on that dimension. Keep in mind that in an experiment, these dimensions are turned into stimuli with colors, shapes, sizes, positions, and so on. Subjects do not get to see 1s and 0s as we do here, and so learning is harder than it often seems from such diagrams.) As can be seen, for the top pair of categories, each member of Category A has three or more 1s, but no member of Category B does. Therefore, a learner could acquire these categories by learning which features (the 1s) tend to go with Category A, and simply deciding whether each item has a majority of those features. In contrast, the category pair at the bottom of the table does not have any such rule. Although Category A does have more 1s than does Category B, exemplar B3 has a majority of 1s, and exemplar A1 has a majority of 0s. There is no way to simply add up the features to decide category membership. Instead, in order to learn this category, one must learn about individual exemplars (memorizing that A1 is a member of Category A, that B1 is in Category B, and so on), or one
Table 6.2.
Category structure used by Wattenmaker et al. (1986).

<table>
<thead>
<tr>
<th>Linearly Separable Categories</th>
<th>Nonlinearly Separable Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category A Exemplar</td>
<td>Category B Exemplar</td>
</tr>
<tr>
<td>A1</td>
<td>1</td>
</tr>
<tr>
<td>A2</td>
<td>1</td>
</tr>
<tr>
<td>A3</td>
<td>1</td>
</tr>
<tr>
<td>A4</td>
<td>0</td>
</tr>
</tbody>
</table>

Note: D1, D2, etc. refer to stimulus dimensions of each item. 1 and 0 refer to the values on those dimensions. Each exemplar is made up of four features, one from each dimension.

must learn about configurations of features (e.g., learning that only members of Category A have a 1 on D2 and a 1 on D4).

Wattenmaker et al. (1986) believed that whether a linearly separable or non-linearly separable category would be easier to learn could depend on how the structure was related to knowledge that people bring to the learning situation. In some cases, the knowledge would suggest that the evidence should simply be summed up—consistent with a linearly separable category. In other situations, the knowledge would refer to configurations of features—consistent with a non-linearly separable category. In their Experiment 1, they used personality features. They hypothesized that people are believed to be kind or intelligent or extraverted to the degree that they evince behaviors consistent with these characteristics. A kind person may not always be kind but is kind more often than not. Thus, personality categories might be particularly susceptible to a linearly separable form of categorization.

In their experiment, Wattenmaker et al. manipulated knowledge by the way they assigned features to their design. In the trait condition, the 1s and 0s in table 6.2 were replaced by actions that were consistent with a given personality. For example,
the 1s on each dimension might be replaced by actions consistent with an honest person, whereas the 0s would be replaced by actions of a dishonest person. Pattern A1 in the linearly separable condition was: "Returned the wallet he had found in the park; Admitted to his neighbor that he had broken his rake; Told the host that he was late for the dinner party because he had overslept; Acted like he enjoyed shopping when his girlfriend asked him to go along with her to the store" (Wattenmaker et al., p. 169). The first three properties (1s) display honesty, whereas the last (a 0) is less than fully honest. Thus, in the trait condition, summing up the properties would lead to a simple rule such as "the people in Category A are usually honest." In the control condition, the dimensions were four different traits (honesty, talkativeness, cooperativeness, and cautiousness). Here, subjects would not be expected to sum up the features, because there is no personality category that corresponds to all these traits. That is, no prior knowledge structure says that people in a given category should be both honest and talkative, but not cooperative or cautious. So, a 1 on D2 was not related to a 1 on D4.

The results fit Wattenmaker et al.'s predictions. In the trait condition, the linearly separable categories were easier to learn. However, in the control condition, subjects found the nonlinearly separable category slightly easier to learn. Presumably, subjects were able to use their knowledge of personality types to combine the different actions in order to use a trait description of the category. In the nonlinearly separable case, this was not possible, and so attempting to use the trait was not very successful. In later work, Wattenmaker (1995) has suggested that social categories in general are interpreted as linearly separable, in part because we do not expect people to be entirely consistent in their behavior. Thus, social categories in general are flexible, in that they do not require specific configurations of features but usually just require some preponderance of evidence ("most of the time, Jessica's quite pleasant").

Wattenmaker et al. (1986) found that it was also possible to induce subjects to prefer a nonlinearly separable category. Without going into the details of their experiment, I'll describe a stimulus to illustrate it (based on their Experiment 3). Suppose that you are trying to learn about an occupation. Half of the members of this group work in the winter and half in the summer. Therefore, the season in which they work does not seem to distinguish this group from any other. However, later you realize that half of these people work indoors and half work outdoors. Now the season does seem relevant: The indoor workers should work in the winter, and the outdoor workers in the summer. This configural information is not the linearly separable sort, because it depends on how the two features are configured—"works indoors" is not predictive of the category by itself; it's predictive when it occurs with
“works in winter” and not predictive otherwise. You can’t simply sum up the evidence. Wattenmaker et al. (1986) found that when pairs of features made sense, as in this example, subjects found the nonlinearly separable categories easier to learn than the linearly separable ones. When the feature pairs did not make sense (i.e., were not predictable based on prior knowledge), then the difficulty of learning reversed.

In short, Wattenmaker et al. showed that which structure was more difficult depended on how it related to prior knowledge. Furthermore, prior knowledge was not always helpful. When it was well matched to the category structure, it did improve learning. But when it did not match that structure, it actually hurt learning in some cases. So, knowledge in and of itself may not be helpful unless it actually conforms to the category’s structure. These results suggest that we should not think of prior knowledge as being an attribute of a category that can be simply added or taken away in the way that a single feature can. The utility of prior knowledge may depend in a subtle way on how the features related to knowledge are empirically structured. This is an important issue on which more work needs to be done.

Category Construction
As described in chapter 5, a very different way to study category acquisition is not to teach people the category but to let them try to discover it on their own—the category construction task. In real life, people notice distinctive classes of objects by themselves without someone telling them that these are different sorts of things, or instructing them in the name. In experiments of this sort, researchers have typically given the subjects a set of stimuli printed on cards and then asked them to divide them into the groups that are best or most meaningful. Quite surprisingly, adult subjects overwhelmingly choose a single stimulus dimension on which to divide up the cards rather than forming family resemblance categories (Ahn and Medin 1992; Medin et al. 1987; Regehr and Brooks 1995). For example, they might divide up the items based only on their color. Given that real-world categories like chair and rose seem to be family resemblance categories, this result is rather puzzling. You can’t identify a rose simply by its color or its having thorns: You need to use many features associated with roses. The reasons for subjects’ strong preference for unidimensional categories are discussed in chapter 5. Here I focus on the influence of knowledge on this task.

Perhaps if subjects had knowledge connecting the features, they might attend to more of the features when constructing these categories. Medin et al. (1987) had used pictures of bugs in their experiments, and they thought that if they impressed subjects with the idea that these were different species of bugs inhabiting different
ecosystems, subjects might look for larger constellations of features rather than choosing just one, such as head shape. In one experiment, subjects were told that the bugs represented bottom dwellers and top dwellers of a pond. They were shown the two category prototypes and told that one was the “best adapted top dweller” and the other the “best adapted bottom dweller.” Even with this help, 19 of 20 subjects used a single dimension to divide up the stimuli. That is, rather than noticing that a bunch of items were similar to one prototype and another bunch of items similar to the other prototype, they might put together all the items with the same head shape as prototype 1 and then all the items with the same head shape as prototype 2, ignoring the other dimensions. So, clearly, very general background knowledge about animals is not sufficient to overrule this unidimensional bias.

Ahn (1990) tried a more direct approach. She used features of flower categories that were clearly different, such as color and location, and provided subjects with a causal explanation connecting the properties of the two kinds of flowers, namely that one type of flower attracted a kind of bird, and the other attracted a kind of bee. The bird was said to like bright colors, be active at night, fly high, and lay eggs near water, and the bee was described as having the opposite values. After hearing this information, subjects divided the stimuli up into two categories. Under these conditions, subjects did form family resemblance categories a third of the time, identifying flowers that were brightly colored, blooming at night, found in trees, and near the water as one category. This result does demonstrate that knowledge can help people to identify category structure. However, the knowledge provided was somewhat arbitrary, in that the bird or bee just happened to prefer the prototypical features of the two categories, and subjects were given these descriptions that connected all the categories' features right before sorting. If, then, this sort of information is required to induce family resemblance categories, it would not be likely to be very helpful in more realistic settings.

Spalding and Murphy (1996) and Kaplan and Murphy (1999) used the stimuli like those described in table 6.1 in category construction experiments. They found that explicit instruction in the connection of the features is not necessary with these items, because subjects could spontaneously identify the relations between features. However, it was important to ensure that subjects first examine all the items. Spalding and Murphy (Experiment 3) found that if subjects were simply given cards and asked to divide them up, they discovered the family resemblance category 40% of the time when the categories were distinguished by themes (like the arctic and jungle vehicle). When subjects first read through all the items and then were asked to divide them up, 78% of the categories were family resemblance sorts. Thus, one
reason why people often make unidimensional sorts in experiments may be that they do not study the items carefully enough to observe whatever structure is there. However, it should be emphasized that when there is no knowledge relating the features, examining the cards by itself is not helpful. In this condition, Spalding and Murphy found that no subject recovered the category structure.

Kaplan and Murphy (1999) essentially repeated their experiment with category learning described above, only using a category construction task. Recall that in their learning experiment, every item had only one feature that was related to the theme. All of the other features were neutral. In spite of the very small number of knowledge-related features in their stimuli, a third of the subjects divided them up into family resemblance categories. Control subjects who did not get these few knowledge-related features never made family resemblance categories. Thus, even a small amount of knowledge can lead subjects to pick out the category structure. One surprising result from their experiment was that when the categories had a thematic basis, subjects not only learned about the thematic features but also learned about the nonthematic features to some degree. Subjects whose categories did not have a thematic basis did not learn about either. The theme helped subjects notice the empirical structure of nonthematic features, which suggests an interaction between the empirical and knowledge-based learning processes.

These experiments demonstrate that knowledge may be helpful in people’s spontaneous creation of categories. As described in chapter 5, the category construction methodology may be somewhat artificial, and so its results should be taken with a grain of salt. Nonetheless, the effects of knowledge demonstrated in these experiments suggest that knowledge may be used in more realistic situations as well. For example, suppose that you notice a tree with strangely shaped leaves that you’ve never seen before. You could make a new category that is essentially unidimensional: tree with such-and-such a leaf. However, your knowledge of plants suggests that other properties are likely to be relevant, such as the location where the tree was found (forest, urban, mountain, near water), its size, the presence or absence of fruit, and so on. If the tree is in an unusual environment (say, a desert), you might try to make a connection between the strange shape of the leaf and the environment. Perhaps the leaves are small, to reduce water loss. In short, it is unlikely that you would stop at one dimension when trying to understand this kind of tree, and the other dimensions you consider would probably be the ones that your knowledge suggested were relevant.

The effect of knowledge in unsupervised category formation is important, because it indicates that feedback is not necessary to obtain the benefits of knowledge. In
many situations one must form concepts without anyone available to provide instruction; in particular, it has been suggested that many of children’s early concepts are initiated in just this way (see chapter 10). The somewhat surprising finding that even small amounts of knowledge can aid in this task (Kaplan and Murphy 1999) provides further evidence that children may benefit from knowledge when learning categories. That is, some have suggested that children do not know very much about biology or social categories or mechanical devices, and so their background knowledge could not help them very much. However, even a small amount of knowledge does improve category construction.

Knowledge in Categorization

The examples discussed up until now have all dealt with the initial acquisition of a category. However, knowledge may also influence a later-occurring process, namely the categorization of items into categories that are already known. An example of Douglas Medin’s (Murphy and Medin 1985) will illustrate. Suppose you were at a party, and you heard that one of the guests had fallen into the pool. You might conclude that this person was drunk. But this is not because you have stored in your concept of drunks the feature “falls into pool.” Instead, this event can be explained by the categorization of the person as drunk. Similarly, there are new inventions and newly seen objects that may not be very similar to anything we know, but which can be explained by their category membership. For example, suppose that you saw a construction worker holding a loud device that seemed to be shooting nails into some wooden studs. You might understand this machine to be a kind of electric hammer, even though it looks very different from other hammers you’ve seen. However, it fills the same function as a hammer and presumably was invented to fulfill the same purpose a hammer was invented for. In short, the machine would be explained if it were a hammer. As another example, consider a difficult medical diagnosis. (Diagnosis is a form of categorization, and patients with the same disease often do not have the same set of symptoms.) In a complex case, a patient may not look exactly like any single disease victim that the doctor is familiar with. However, perhaps the patient’s symptoms can be explained by a particular set of disease categories (e.g., atypical hepatitis combined with a vitamin E deficiency). Doctors can make such diagnoses through reasoning processes that tell them that these diseases could explain the patient’s symptoms and test results. All of these examples concern not the original learning of a category but rather the process of deciding whether something is or is not a category member.
In short, categorization may not always be based on simple matching of properties. It may be that background knowledge is more actively involved. There are many anecdotes of the sort just described, but there have been fewer tests of this hypothesis than in category learning. Most of these tests have been with natural categories, since these are ones for which subjects are most likely to have well-developed knowledge structures.

One study that examined categorization of this sort was performed by Rips (1989), who created a scenario in which an object or animal went through a transformation. In particular, he described to subjects something that was much like a bird. This animal, however, had the misfortune to live next to a toxic waste dump, where chemicals caused it to lose its feathers and develop transparent, thin wings. It also grew an outer shell and more legs. At the end of this metamorphosis, the creature appeared rather reminiscent of an insect. Rips asked subjects whether the animal was a bird or an insect and whether it was more similar to birds or insects. Subjects claimed that this stimulus was more similar to an insect but that it was still more likely to be a bird. That is, subjects felt that the change in outer appearance and even body parts did not constitute identity, but that the animal’s inherent biological properties determined what it was. Apparently, then, this judgment was based on a domain theory about biology rather than on more superficial properties.

Keil (1989) performed some very similar studies on children and adults using a surgical scenario, described in chapter 10. He found similarly that adults were not persuaded that transforming an animal would make a difference in its category membership, even when it ended up looking exactly like a different item. Children were somewhat more flexible: Before grade 2, they were generally willing to believe that a cat could be turned into a skunk, for example. However, even children at this age did not believe that an animal could be transformed into a plant or inanimate object.

Rips (1989) reported another influential study using a different kind of example. Imagine that there is a round object that is half way in size between a quarter (a U.S. 25-cent piece) and a pizza. Is it more likely to be a quarter or a pizza? There does not seem to be any way to tell based on empirical evidence. However, because people realize that coins are restricted in size by convention and law, whereas pizzas are not, Rips expected subjects to categorize it as a pizza. And in fact, this is what most subjects did.

Unfortunately, this study was found to be difficult to replicate, as Smith and Sloman (1994) discovered that the choice of the nonvariable category was found only when subjects were required to talk out loud and justify their answers. If they
Figure 6.2
Pictures of the tuk category used by Lin and Murphy (1997). Subjects read descriptions of the item and were told what each of the numbered parts was. However, different subjects received different information about the objects and their parts, as explained in the text.

only had to make a categorization decision, they chose the variable and nonvariable categories about equally. Although this still indicates that people will use their background knowledge in making these decisions, Smith and Sloman argued that people use such knowledge only when encouraged to do so by the instructions of the task. In everyday life, they suggest, categorization of entities may not involve such knowledge at all, especially when it is done rapidly without much overt reflection.

Lin and Murphy (1997) attempted to find evidence for more automatic, less reflective use of knowledge, with speeded identification tasks that have been used in past categorization research (e.g., Murphy and Brownell 1985; Smith, Balzano and Walker 1978). These tasks would not be of the sort Smith and Sloman criticized, since they require fast, nonverbal responses. Subjects first learned about kinds of artifacts in a foreign country, such as the set shown in figure 6.2. Different subjects were taught different things about these items. For example, one group was told that the tuk displayed in figure 6.2 was used for hunting. The hunter would slip the noose (1) at the top over the animal’s head, and pull on the loose end of the rope (4) to tighten it. The hunter held the tuk at the handle (3), and the hand guard (2) protected the hunter from animal bites and scratches. A different group read a very different story about the tuk. They were told that it was a fertilizing tool. The liquid
fertilizer was held in the tank (2), and the knob (3) was turned in order to let it flow out the outlet pipe (4). The loop at the top (1) was used to hang up the tuk in storage. Obviously, the two groups have very different background knowledge about this implement. However, they have both seen the same examples of tuks and have been given descriptions of the same parts. The question is, then, whether this difference in background knowledge will influence subjects' categorizations.

Lin and Murphy tested this question by creating items that lacked one or more of the parts of the originally seen tuks. As shown in figure 6.3, one item might lack the loop at the top, one might lack the triangular part in the middle, and another might lack both. The descriptions of the objects had been designed so that the most important part in one description should be relatively unimportant in the other description. For example, the loop at the top is critical for the hunting tool tuk, because it grabs the animal. In contrast, that part is not essential for the fertilizing tuk, because it only concerns the storage of the tool. The reverse is true of the triangular part (2), which is essential for holding the fertilizer but is only a convenience for the hunting
tool. Lin and Murphy found that subjects' willingness to call these items *tucks* depended on the initial description. Those who learned the tuk as a hunting tool did not categorize the item with the missing loop at the top as a tuk; those who learned the item as a fertilizer did. These effects were also found in reaction times of speeded category decisions, even when subjects were given a response deadline, to make them go faster than usual. In fact, similar effects were found for subjects whose task was simply to decide whether the object, when presented briefly, had all of its original parts or not: Subjects were more likely to notice when a functionally critical part was missing than when a less important part was missing.

In short, although all subjects saw the same examples, they drew different conclusions about what parts of the object were most important. These conclusions influenced speeded categorization decisions, even though this background knowledge was not specifically asked about during the task. Palmeri and Blalock (2000) also found knowledge effects in speeded categorization, even when subjects were forced to respond very quickly. This result gives some confidence that Smith and Sloman's (1994) finding of limited knowledge effects is not universal. Recall that Smith and Sloman found that subjects only seemed to use their background knowledge when they were asked to justify their answers. In contrast, in Lin and Murphy's and Palmeri and Blalock's experiments, the results were found without such justifications being asked for, and under speeded instructions. (Especially noteworthy is the finding of knowledge effects on part detection, which is a simple, factual question.) Thus, some kinds of knowledge are probably directly incorporated into the category representation and used in normal, fast decisions about objects. Other kinds of knowledge, however, may come into play only when it has been solicited.

More generally, it seems unlikely that much background knowledge is called for when you walk down the street and see a pigeon, say, on the sidewalk. If the pigeon is clearly visible, and you've seen many of them before, you will identify it extremely quickly as a bird, without having to go through a long reasoning process. In other circumstances, however, overt reasoning using background knowledge would be called for. As described earlier, cases of medical diagnosis often require one to reason explicitly about how to explain the symptoms and test results observed. More generally, if you have to categorize something based on minimal perceptual information (e.g., a noise at the window), you might naturally rely almost entirely on background knowledge to arrive at an answer (a bird might have flown up to the window and then flown away; a dog couldn't get up there and then disappear so suddenly). In between these two extremes are cases in which knowledge might be somewhat involved or could speed up or slow down a decision. For example, if you were in the desert, you might have trouble identifying something on the ground as a pigeon,
Table 6.3.

<table>
<thead>
<tr>
<th>Mornek Category</th>
<th>Frequency</th>
<th>Plapel Category</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Installed near garbage dumps*</td>
<td>.50</td>
<td>Located in a nuclear plant*</td>
<td>.50</td>
</tr>
<tr>
<td>Found near mosquito-infested swamps*</td>
<td>.50</td>
<td>Found in the city water supply*</td>
<td>.50</td>
</tr>
<tr>
<td>Produces a poisonous substance*</td>
<td>.50</td>
<td>Has a red flashing light*</td>
<td>.50</td>
</tr>
<tr>
<td>Emits microwaves*</td>
<td>.50</td>
<td>Makes a loud beeping noise*</td>
<td>.50</td>
</tr>
<tr>
<td>Turned on by a dial</td>
<td>1.0</td>
<td>Turned on by a key</td>
<td>1.0</td>
</tr>
<tr>
<td>Box shaped</td>
<td>1.0</td>
<td>Barrel shaped</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Note: Asterisks indicate knowledge-relevant features. Frequencies are the proportions of category items containing that feature.

because you would not expect pigeons to be there. Little is known about these in-between cases and how much knowledge is involved in the sort of everyday object categorization that people do so effortlessly.

Wisniewski (1995) addressed part of this issue by asking whether empirical information or background knowledge is more important in categorizing things when the two conflict. He constructed categories using features shown in Table 6.3. In this example, subjects learned about two novel tools, morneks and plapels. Morneks tended to have the features listed in the left column, whereas plapels tended to have the features listed on the right. (Individual examples were made up by sampling from these features, plus some random features that occurred equally often in both categories.) Also, some features were quite typical of their category: They appeared in every example, as indicated by the 1.0 in the Frequency column. Other features occurred in only half the category items, as indicated by .50 in the Frequency column. The knowledge group was told that morneks are used for killing bugs, and that plapels are used for detecting toxic substances. The ignorance group (a name I have invented to emphasize the difference between conditions) was not given this information. This information about the function, then, is the background knowledge that Wisniewski manipulated. After learning, subjects had to categorize test items and say how confident they were in the category decision. If people use their background knowledge to categorize items, then there should be a difference between these two groups’ performance on test items.

Some of the test items used are shown in Table 6.4. The first type of item, the atypical test item pits the function against the statistical structure of the category. Two of the features (listed first here) are strongly related to the object’s function:
Table 6.4.
Examples of Test Items for the Mornek Category (see table 6.3) used by Wisniewski (1995).

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Functional Relevance</th>
<th>Feature Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Atypical Test Item</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emits microwaves</td>
<td>relevant to Mornek</td>
<td>.50</td>
</tr>
<tr>
<td>Installed near garbage dumps</td>
<td>relevant to Mornek</td>
<td>.50</td>
</tr>
<tr>
<td>Turned on by a key</td>
<td>irrelevant</td>
<td>0</td>
</tr>
<tr>
<td>Barrel shaped</td>
<td>irrelevant</td>
<td>0</td>
</tr>
<tr>
<td>Inspected once every 6 months</td>
<td>irrelevant</td>
<td>.50</td>
</tr>
<tr>
<td><strong>Function Test Item</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emits microwaves</td>
<td>relevant to Mornek</td>
<td>.50</td>
</tr>
<tr>
<td>Installed near garbage dumps</td>
<td>relevant to Mornek</td>
<td>.50</td>
</tr>
<tr>
<td>Inspected once every year</td>
<td>irrelevant</td>
<td>.50</td>
</tr>
<tr>
<td>Operates during the night</td>
<td>irrelevant</td>
<td>.50</td>
</tr>
<tr>
<td><strong>Nonfunction Test Items</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Turned on by a dial</td>
<td>irrelevant</td>
<td>.50</td>
</tr>
<tr>
<td>Box shaped</td>
<td>irrelevant</td>
<td>.50</td>
</tr>
<tr>
<td>Inspected once every 6 months</td>
<td>irrelevant</td>
<td>.50</td>
</tr>
<tr>
<td>Operates during the day</td>
<td>irrelevant</td>
<td>.50</td>
</tr>
</tbody>
</table>

Emitting microwaves is a possible way of killing bugs, and garbage dumps are certainly a location where a bug-killer might be useful. The knowledge one might bring to bear on this item would support its being a mornek, the bug killer. However, two other features (listed next) were strongly associated with the pnapel category. As table 6.3 shows, “turned on by a key” and “barrel shaped” occurred in all of the pnapels and in none of the mornekas. Thus, this is a very strong statistical clue to category membership. (The final feature was equally present in both categories.) Subjects in the ignorance group put this item into the mornek category only 43% of the time. However, subjects in the knowledge group were apparently more swayed by the relevant features, and they placed it into the mornek category 65% of the time. When two features were related to knowledge—that is, were closely related to the function of one category—they overruled the two features that statistically predicted a different category.

This contrast is also made by the two test items underneath. Here, the statistical properties of the features were held constant—all of them occurred in half of the learning items. However, in the function item, two properties were related to the object’s function, but in the nonfunction item, none of the features was. These two differ, then, in the degree to which they relate to the knowledge underlying the
category. Because these items are unambiguous, subjects almost always categorized them correctly. But Wisniewski found differences in how confident the subjects were about their answers. The knowledge group was more confident about the function items than about the nonfunction items, but the ignorance group showed no such difference (see also Wisniewski 1995, Experiment 1). Thus, the knowledge that subjects were given about the function influenced their categorization both for somewhat unusual exemplars (the atypical test item) and typical ones.

**Locus of the Knowledge Effect in Categorization**

There are two general ways that knowledge could be having its effect on categorization. First, when the knowledge is present during learning, it could be influencing the nature of the category representation. For example, Lin and Murphy (1997) proposed that when people learned their tools (see figure 6.2), they might have paid more attention to the features that were critical to the function. When the tuk was used for catching animals, they might have encoded more information about the loop at the top and elaborated their memory with inferences about how the loop worked, and so on. In contrast, when the tuk was used for fertilizing crops, this feature would not have received much attention and would have been represented in less detail. Second, it is possible that the knowledge is activated and used during the categorization judgment itself, after the learning period has ended. For example, at test, subjects in this experiment might have thought something like “that thing couldn’t catch an animal, because it doesn’t have the rope at the end, so it can’t be a tuk.” Here, it is not just that the initial representation of the category has been influenced—the underlying knowledge itself is actively used in making the categorization decision. Of course, it is possible that knowledge is used in both ways.

The criticism of knowledge effects by Smith and Sloman (1994) is about the second sort of use. It seems implausible to some researchers that when people see an object, they activate rather complex knowledge structures and then draw inferences about the object’s identity. Object categorization is often quite fast (for familiar categories at the basic level, anyway), usually much less than a second, and it seems unlikely to some that much knowledge could be activated and used in that time. In experiments with time constraints on responding (Lin and Murphy 1997; Palmeri and Blalock 2000), it seems even less likely that the explicit use of knowledge is involved. Furthermore, Lin and Murphy found effects of background knowledge even in a task where functional importance is not relevant, namely part detection. Thus, their effects are probably due to knowledge influencing the categories’ (and objects’) encoding in memory.
Wisniewski (1995, Experiment 3) investigated this issue with regard to his categories by comparing groups that received the critical knowledge either before or after learning. The early knowledge group was given the information about the object's function during the learning trials, as in his other experiments. The late knowledge group went through the learning trials and was only told about the functions associated with the categories at their end. So, they could not have used this information to encode the properties of individual objects during learning, but they could still use this information in making their categorizations during the test phase. In fact, Wisniewski found that both groups seemed to use this knowledge. For example, both groups rated the function test items higher than the nonfunction test items, as described above (and see table 6.3). However, such differences were significantly larger for the early knowledge group. Knowledge was actively used during the categorization phase, as shown by the late knowledge group. But knowledge also influenced the learning of the exemplars, as shown by the greater use of knowledge for the group that received it prior to learning. Wisniewski's stimuli were verbal lists, so whether such results would extend to objects has yet to be seen.

Before the possible sources of knowledge effects in categorization can be fully understood, we will need more information about when and where such effects occur. It seems more likely that knowledge will actively be used during categorization when the decision is difficult, is slow, is based on little perceptual information, and in similarly straitened situations. It may be that the use of knowledge varies considerably both with category and with the pressures on the categorization task. When knowledge influences the initial encoding of a concept, however, that knowledge has been incorporated into whatever representation is used to later categorize items, and the categorization process cannot help but be affected by it.

A somewhat different case that might well be taken as demonstrating the use of knowledge in categorization is Barsalou's (1983, 1985) study of ad hoc categories. These are categories that are constructed in specific situations to describe a specialized class of objects that are of particular interest (see Barsalou 1991, for more detail). Examples include things to take on a camping trip, things to carry out of a burning house, ways to avoid being killed by the mafia, and ways to make new friends. Although these categories are of interest in a number of respects, for our purposes, the most important thing is that people can verify objects as being members of these categories, as well as being more or less typical members. For example, which of the following do you think are good examples of ways to avoid being killed by the mafia?
move to South America
have a garage sale
take a bath only once a month
change where you’re living in Las Vegas
change your identity

Most people would agree that the first and last items are both good examples of this category. “Change where you’re living in Las Vegas” is a borderline member in my estimate. It is somewhat similar to “move to South America,” but it would be unlikely to help you avoid the mafia for very long. The interesting point about these categories is that people can make such judgments with high reliability (Barsalou 1983) even if they have not previously learned the category or thought about the items in this way. I doubt that you learned item by item whether the things on this list were examples of ways to avoid being killed by the mafia. For most ad hoc categories, there is no learning period, per se. Instead, the phrase evokes a certain kind of meaning, which you can then apply to a new item. By reasoning about whether the new item fits this meaning, you can decide whether it’s a category member or not. So, taking a bath only once a month may be off-putting, but you probably cannot think of a reason that it would deter a hardened mafia hit man who has been instructed to kill you. In contrast, by moving to South America, you might be out of range of the mafia’s ability to find or harm you. These are not facts that you learned previously, but inferences you made at the time of decision. In short, category decisions about ad hoc categories are further evidence for the use of knowledge during categorization itself. Because these categories usually do not have any previous learning phase, the only way that knowledge could be influencing judgments is during the categorization process.

What clearly can be concluded from this discussion is that background knowledge affects not only initial acquisition of a concept but also later categorization judgments. People tend to positively categorize items that are consistent with their knowledge and to exclude items that are inconsistent, sometimes even overriding purely empirical sources of information. The details of how this works have not yet been established, largely because few experimental studies have been done, and most have not used objects but instead have employed verbal descriptions. Also, it seems very likely that how and when knowledge influences categorization will depend on the nature of the category and whether other cues are readily available. Working out the mechanisms of these processes is one of the more important tasks for this approach to concepts.
Knowledge in Feature Construction

Throughout this book, I have been talking about objects and concepts having certain features. For example, birds fly and have two legs, robins have red breasts, my pet bird Tweety sings, and so on. The reader may have wondered where these features come from—that is, not where do robins or their legs come from, but how is it decided that these particular properties are the ones that are part of the robin concept, as opposed to other properties? It is not too controversial to say that any object could be thought of as having an infinite number of features. For example, robins have two legs, but they also have a leg; they also have fewer than three legs, and fewer than four legs, and so on. Robins have a red breast; or, one could say that they have a red chest and red belly. Which, if any of these, should be features? Why is it that red breast gets to be the feature, rather than red chest plus red belly? Why isn’t having fewer than four legs a feature? And I saw a robin on my lawn this morning. Should this fact (“found on my lawn this morning”) become a feature of robins? Or should the property “robins can be found on lawns” or “robins can be seen in the morning” or “robins can be found in the morning in the United States” … or … “in the Midwest” or “in Illinois”? Any object or event can be conceived of in many different ways, and it would be impossible to encode each one and store it as part of a concept.

We do not know very much, unfortunately, about which properties are encoded, or how this is determined. However, it has been argued that knowledge might play an important role in this process (Goodman 1965; Murphy and Medin 1985). One’s prior knowledge of a domain provides a set of properties that can be used in encoding a new member of that domain. Indeed, it seems likely that in learning a new category, you compare it to similar categories and use features from or based on those categories to construct the properties by which to represent the category (Markman and Wusniewski 1997). This kind of comparison may not be a “knowledge effect,” per se; it may only involve contrasting a new category to an old one from the same domain.

Schyns and colleagues (Schyns, Goldstone, and Thibaut 1998; Schyns and Murphy 1994; Schyns and Rodet 1997) have argued that features are in part defined by how they are related to category structure. In some cases, the features are so ambiguous that they cannot be reliably identified on simply examining a stimulus. Consider, for example, an X-ray. Often X-rays have a large number of spots and shadows on them that are of questionable importance. It is only experts who know that some spots are uninteresting structures (“You always get these shadows when
you photograph through the ribs."), or artifacts of the procedure ("She probably moved a little while the X-ray was taken."). whereas others have special significance ("This series of spots could be an unusual growth."). When novices look at X-rays, they simply do not know which markings constitute coherent features, and which do not (Lesgold 1984). In these cases, category learning itself may be needed to identify the features. That is, learning to diagnose conditions from an X-ray does not just require learning which features predict which diseases but also requires learning which patterns are features and which are not.

There is evidence for this idea in the work of Schyns and Murphy (1994). They used blob-like stimuli with very ambiguous parts, such that subjects could not identify the critical features in advance. That is, they couldn't accurately determine which sections of a blob were a part. Subjects learned to distinguish two categories of such blobs. At the end of learning, they were now able to identify the correct features within the blobs, namely, the features that distinguished the categories, because they had learned that some aspects of the blobs were found in one category and some were found in the other. Control subjects who simply viewed the stimuli were not able to identify the same features—category learning itself was required. One might worry that the task, in which subjects were explicitly asked about the parts of the objects, might have been responsible for the results. Schyns and Rodet (1997) addressed this with a clever design in which they taught subjects categories in sequence. They proposed that subjects who formed features in the course of learning the first categories would find it easier to learn a second set that involved the same features. Here no explicit identification of parts or properties was required. The results supported their hypothesis.

In many situations, we assume that the features are obvious, and the only difficulty subjects have is in learning which features are associated with which categories. These experiments suggest that in some cases, the features are not readily available at the beginning of learning and that one goal of learning is to acquire the features themselves.

These examples, though, are not very knowledge-dependent in the way we have defined "knowledge" in this chapter. It was not general facts that people knew about the domains that determined the feature learning here, but instead the learning of the particular category that accounted for it. A more radical demonstration of knowledge-based feature creation was carried out by Wisniewski and Medin (1994). In their study, the stimuli were children's drawings of people. They divided up the pictures into two categories. One category had detailed drawings that also shared some superficial properties, such as curly hair on the people. The other category had
pictures of people performing some action; they also shared some superficial properties such as not smiling. In one experiment (1a), subjects were shown these two collections of pictures and were asked to come up with a rule that separated the two categories. Subjects were told either that one group of drawings was done by creative children and the other by noncreative children, or else that they were drawn by children in Group 1 and Group 2. The use of these meaningful category names in one condition was a way of activating knowledge structures that might be used in interpreting the pictures.

Wisniewski and Medin examined the features that people produced as part of their rules. They found that about half of the features listed by the subjects who did not have prior knowledge (the Group 1/Group 2 subjects) were simple, concrete features that were readily detectable in the pictures. One such rule was “all of the characters have their arms straight out from their bodies, and they’re also standing very straight” (p. 241). In contrast, only 12% of the features of the knowledge group were so simple. Instead, their features were more likely to be abstract, describing a whole class of properties (e.g., “drawings that show more positive emotional expression”). In addition, the knowledge subjects were likely to make hierarchical features, features that had both an abstract and a concrete component. In these cases, the subjects would mention an abstract property of the features and then relate it to a concrete feature: “There are many more details in each of them [drawings], things like belts, pockets, and patterns on the clothes...” Here, the abstract feature of “details” is mentioned and is related to more concrete properties.

Why did the knowledge group create these abstract and hierarchical features? Wisniewski and Medin suggest that the prior knowledge created certain general expectations about the properties. For example, subjects might expect that creative children would make more detailed or unusual drawings. However, as discussed at the beginning of this chapter, it is difficult to specify in advance exactly how these general expectations would be instantiated. Unless you know a lot about children’s drawings, you might not immediately know whether one was particularly unusual or detailed. Part of subjects’ learning therefore was to relate the features observed in the pictures to these general expectations (i.e., a post-hoc use of knowledge, as described earlier). This process resulted in more hierarchical features being constructed, in which abstract expectations were linked to observed features. In contrast, when subjects did not have any particular expectation about the categories, they were more likely to focus on concrete features, because they did not have an abstract notion in mind at the beginning.

In their Experiment 2, Wisniewski and Medin (1994) asked subjects to provide a rule to categorize the items after seeing each exemplar. Subjects would study a pic-
ture, say which category they thought it was in and why, and then get feedback on the answer. Subjects were told either that the pictures were done by farm kids and city kids or by creative kids and noncreative kids. Because subjects' initial guesses about the items were not particularly accurate, they were often forced to re-evaluate their rules and the features involved. Wisniewski and Medin identified a number of processes that subjects followed in making these revisions. One example was that subjects often reinterpreted a feature as supporting a different category. For example, one subject felt that the "perfect body proportionment" of a drawing indicated a creative child. When told that this was a drawing of an uncreative child, the subject decided that "perfect body proportions show lack of imagination." So, the same feature could be related to a different concept, via a different reasoning process. Another strategy that people used was to focus on new features that might overrule the feature that had led them astray. This is the kind of process that statistical theories of learning would propose—forming associations to the correct features, while weakening associations to other features. However, some of the processes were much more knowledge-related. For example, sometimes subjects changed their criterion for identifying a feature. Although a picture had seemed detailed, when told that its creator was not creative, a subject decided that "drawings done by creative children would be more detailed." Thus, the identification of the feature "detailed" changed as a result of learning. Finally, sometimes subjects reinterpreted the same property based on feedback. For example, one subject identified the clothing in a drawing as a "city uniform," but changed this to a "farm uniform," when told the drawing was done by a farm child. So, the feature itself changed as a result of new information.

These strategies suggest that subjects were not working from a pre-selected set of features that were obvious from the stimuli themselves. Different people focused on different features, and exactly what counted as a feature was rather flexible. Although a picture might seem detailed at first, later experience could change it to being thought of as not detailed. In fact, some subjects looked at exactly the same things and gave different interpretations. One drawing contained a series of dots on the front of the person's shirt. A subject who thought the drawing represented a creative child interpreted this as "buttons," reflecting a detailed picture. Another subject who thought it might be a city child's drawing interpreted the same dots as being a tie. By starting from different expectations, the subjects extracted different features from the pictures.

Wisniewski and Medin (1994) argued that the analysis of an item into features is a function of the stimulus itself, prior knowledge, and learning of the category. People do not view items in a vacuum—they have strong expectations about what
the features will be. However, those expectations are flexible, and they change with actual experience with the category. Thus bottom-up and top-down processes combine to determine what features are associated with a category. And when this occurs, the feature may not be represented as a simple property, like “large” or “blue,” but it may be a more complicated knowledge structure, such as “detailed, therefore contains buttons, pockets and shoes.”

Wisniewski and Medin’s article contains a number of arguments that challenge the usual assumptions by which category learning works. They propose a much more complex interaction of identifying features and learning the category than is normally proposed. Indeed, although their article has been published for a few years, the rest of the field has been somewhat slow to pick up on it, in part because the complexity of the process they describe. It is difficult to model concept learning when the features themselves are changing over the course of learning. Furthermore, one could criticize their study on the basis of ecological validity. To begin with, their categories did not actually correspond to pictures done by creative/noncreative or city/farm children (though their stimuli were real children’s pictures and so were naturally occurring entities, unlike almost all the other experimental stimuli discussed in this chapter). There might, therefore, be some question as to whether the same results would be found when the categories were more accurate. In those cases, how flexible would the features have to be, and how often would they be revised during learning? Also, the learning situations that Wisniewski and Medin used were somewhat unusual, at least relative to learning tasks done in most concept experiments. In their Experiment 1, subjects looked at all the items (simultaneously) and then provided a rule for the categories. In Experiment 2, subjects were required not just to learn the categories, but to provide a rule after every example. These procedures are quite different from the usual categorize-plus-feedback experiment. These are valid concerns, but it would also be a mistake to take the typical category-learning experiment as a standard, simply because it is familiar to us by repetition. In the usual experiment, the stimuli are extremely simple (dot patterns or geometric figures), no information is given about an item other than its category membership, people go through example after example until they perfectly identify the categories, and so on. These characteristics are not particularly representative of real-world learning. Although these techniques have proved useful, it should be kept in mind that all experiments are to some degree simplifications, and the fact that Wisniewski and Medin’s study has a different set of simplifications than usual is not in and of itself a criticism. Nonetheless, it would certainly be useful to have a replication of this work in other kinds of tasks, with other stimuli, and so on. As with
any such ambitious project, it is necessary to replicate and extend it to new situations before its importance can be fully understood.

Since the above was written, an important step in this direction has been taken by Palmeri and Blalock (2000). Although they used Wisniewski and Medin's stimuli, they instituted a more familiar learning procedure and used speeded categorization judgments as their dependent measure. Half their subjects learned two categories of drawings as Group 1 and Group 2, and half learned them as drawings by creative and uncreative children. All subjects were tested on new pictures that matched the old ones either in terms of concrete features (like curly hair) or abstract features (detailed). The results suggested that the neutral category names led subjects to learn concrete features of the pictures, whereas the names referring to creativity led subjects to learn more abstract features. Perhaps surprisingly, these results held even when subjects were forced to categorize the pictures very quickly—within 500 ms. As discussed in the section on categorization, these results suggest that the knowledge influenced the way the pictures were encoded, as Wisniewski and Medin would have expected, rather than a slow reasoning process that operated at the time of categorization. Further work of this sort, especially using new materials, would be welcome.

To sum up, then, most work on concept learning has not said very much about where the object's features come from. By focusing on very simplified stimuli with perceptually obvious properties, this question has been largely avoided. But in real life, the relevant properties of an object or situation may have to be learned in addition to the concepts themselves. An anecdote may illustrate this (from Clark and Clark 1977, p. 486). A young boy misbehaved in some way, and his mother told him, "Young man, you did that on purpose." Later, he was asked what "on purpose" meant, and he replied "It means you're looking at me." Since the boy was told this when he was caught doing something, he inferred that the phrase referred to getting caught, rather than to the much more abstract feature of mental intention that actually determines purposeful actions. Children have to learn this kind of feature just as much as they have to learn the entire category. Learning the relevant properties for mental states is a very difficult task (Wellman 1990), in part because they are not visible. But learning the relevant properties for everyday objects and events may sometimes be difficult as well.

The work I have reviewed suggests that in learning one category it is very useful to already know the features of related categories. In other cases, the features are acquired as a function of category feedback—learning which items are in the category involves learning the features peculiar to that category. In still other cases, the
features themselves may not be simple entities but are knowledge structures themselves, in which general knowledge is related to specific properties. This possibility, which is perhaps the most interesting one, needs further exploration and analysis.

Knowledge in Induction

As described in chapter 8, one of the main functions of categories is induction. Once you know that something is a dog, you have a good idea that it will bark, it is not something you sit on, it has legs, it is male or female, it has a liver, and so on. In their simplest form, such inductions can be read off the concept representation. If you have represented the concept of dog as having the features: barks, has four legs, eats meat, is a mammal, and so forth, then when someone tells you that they have a dog, you can immediately infer that their pet barks, has four legs, eats meat, and is a mammal. This sort of induction does not require any particular knowledge other than the properties associated with the category. However, knowledge can be involved in induction in a more significant way. Because this work is covered in some detail in chapter 8, I will only review it briefly here.

In the standard category-based induction task, one is told that one category (or item) has a given property and then is asked whether the property would be likely to apply to another category (or item). So, the question is about the projection of a novel predicate from one category to another. Researchers have focused on two determinants of this process: the similarity of the two categories and their typicality to a more general category (Rips 1975). Knowledge relating the categories or the property to the categories is not part of the traditional view. Nonetheless, knowledge does influence this process.

Kalish and Gelman (1992) examined cases in which items were in two categories at once, such as wooden pillows or fur bowls. They asked children one of two questions about these kinds of items, which were presented as line drawings as well as being verbally labeled. One question required children to respond on the basis of the kind of item it was (pillow or bowl), and one required children to respond on the basis of the material of the item (wooden or fur). For example, children might be asked to decide which items were soft. Here they should pay attention to the material of the item. In another question, they might be asked whether an object should go into the bedroom or kitchen. Here, they should attend to the kind category. And in fact, children as young as 3 years old responded correctly for both kinds of questions. That is, they said that wooden pillows should go into the bedroom but were hard. Thus, the children knew which categories controlled which features
(showing a very nice sense of conceptual combination to boot). Even though pillows are normally soft, wooden pillows are hard.

How did subjects know to use one category to make one kind of induction and a different category to make the other kind of induction for the same item? Kalish and Gelman proposed (pp. 1555–1556) that children were using knowledge of domains to tell them which features were critical to that domain. The reason pillows are hard or soft is because of the material they are made out of, and wood is by its nature a hard material. So, although pillows are normally soft, the domain of materials controls this judgment. In contrast, the functions of objects determine where they are placed in the home, and pillows are largely determined by their function. Thus, even wooden pillows would be placed in the bedroom.

Ross and Murphy (1999) compared inductions of different kinds of features from the same categories (see also Heit and Rubinstein 1994, described in chapter 8). They compared taxonomic categories of foods, such as fruits, to script-based categories, such as breakfast foods, which are determined by the time, location, or setting in which they are eaten. To do this, they constructed triplets of items: a target food, a taxonomic alternative, and a script alternative. For example, if the target food was cereal, then the taxonomic alternative might be noodles (both are breads and grains), and the script alternative might be milk (both cereal and milk are breakfast foods, but are not in the same taxonomic categories). Subjects made a forced-choice induction judgment. They were told that the target food had a certain property and then had to decide which of the alternatives was more likely to have that property. Ross and Murphy found that when the property was a biochemical one (possession of a given enzyme), subjects chose the taxonomic alternative 83% of the time; when the property was situational (when the food was eaten in a novel culture), subjects chose the taxonomic alternative only 29% of the time. That is, subjects preferred to draw inductions based on taxonomic relations for biochemical properties but based on script relations for situational properties.

This kind of result is a puzzle for most theories of induction, which claim that induction should depend primarily on the similarity of the categories involved. If that were correct, the type of property would not matter, yet these studies show that it does. Ross and Murphy's (1999) results suggest that subjects are engaged in complex reasoning, in which they consider not just how similar the categories are but exactly how they are related and how that relation in turn is relevant to the property being projected. The induction is considered strong to the degree that subjects can create a story that connects the property in the target category to the projected category (see also Lassaline 1996; Lin and Murphy 2000, Experiment 9). An
even more radical effect has been found by Proffitt, Coley, and Medin (2000), who gave induction problems involving disease susceptibility to tree experts. For example, if oaks get a certain disease, would birches also be expected to get it? (They asked the question in a number of different ways, which I'm not distinguishing.) Proffitt et al. found that their subjects often did not simply rely on the similarity of the categories involved but would engage in long chains of reasoning about how a disease could or could not be transmitted from one type of tree to another. They referred to presence in the same ecological setting as well as to specific knowledge, such as how thick the bark is of different kinds of trees. These chains of reasoning went far beyond the similarity relations proposed by the most popular (knowledge-free) models of category-based induction.

All these examples, as well as others described in chapter 8, suggest that prior knowledge and reasoning are heavily involved in category-based induction. This fact has been hidden in most work in the field because most studies look at the induction of blank predicates that are chosen to be as uninformative as possible, often using fictitious or unfamiliar properties. In such cases, subjects can only rely on the overall similarity of categories and similar structural variables. But when the properties make contact with what one knows (as would almost always be the case in real life), people apparently use that knowledge to reason about whether the property should be projected.

Discussion of Models

This chapter has documented a number of ways that prior knowledge influences concepts. (And other chapters have reviewed knowledge effects in the context of their own specific topics, such as conceptual combination, conceptual development, and word meaning.) Indeed, knowledge appears to influence concepts at all stages that have been investigated: in identifying and constructing features that form the conceptual representations, at initial acquisition (in both supervised learning and unsupervised formation tasks), in categorizing novel examples, and in using concepts to make inductions. Later chapters will extend this list to include conceptual development and conceptual combination. These phenomena are consistent enough that a complete theory of concepts will have to explain them, and the effects are pervasive enough to suggest that it will not be possible simply to tack some knowledge onto a purely empirical theory. The knowledge will have to be integrated into the theory's processes of learning and use at various levels. That said, it is not clear that all the phenomena that are here being called "knowledge effects" form a single,
coherent class. These effects may reflect knowledge of different sorts, may involve different interactions of knowledge with the other processes, and may have different constraints on how knowledge is used. Whether all these effects can eventually be incorporated into a single model is an empirical question, but one that we are still far from answering.

As usual, I will now discuss the three main theories’ approaches to these questions. As will be seen, there is still much left to be answered as to how any of these theories will explain the results.

Prototype and Exemplar Models
As they are normally presented, neither prototype nor exemplar models have attempted to account for the knowledge effects described above. The problem is that these models start from a kind of tabula rasa representation, and concept representations are built up solely by experience with exemplars. Most of their attention has been directed toward empirical typicality and learning effects.

For example, in the prototype model (see chapter 3), subjects learn which features are associated to which categories by keeping a rough count of how often each one appears with category members. Features that are distinctive to a given category receive a higher weight than do features that are found in many categories. There is little need for prior knowledge in this process, as the features are assumed to be given directly by the stimuli, and knowledge is not necessary to keep track of the feature count. Similarly, in categorization, an item is compared to the feature list, resulting in a measure of how closely it matches the category representation. Knowledge is not used here, either, as the comparison process requires only the count of matching and mismatching features (Smith and Medin 1981).

Thus, the standard prototype model has no particular need for knowledge to account for the typicality effects that led to its creation, and so it has not incorporated knowledge as part of its learning and categorization mechanisms. The problem, then, is that the model cannot account for the data presented in this chapter that show that knowledge is in fact used. There is nothing in the prototype model, for example, that says that the shape of the whingelow is more important than its present location in identifying what kind it is. It does not provide any way by which features can be constructed, rather than simply observed (Schyns et al. 1998; Wisniewski and Medin 1994). There is no mechanism by which it can explain why categories related by a theme are easier to learn or construct than those that are not (Murphy and Allopenna 1994; Pazzani 1991; Spalding and Murphy 1996; Wattenmaker et al. 1986). In categorization, it does not have any mechanism by which to
explain why some properties are more important than others that were equally often present in the learning set (Lin and Murphy 1997; Palmert and Blalock 2000; Wissniewski 1995). It has no way whatsoever to explain how ad hoc categories are formed, especially since they do not follow the normal rules of family resemblance (Barsalou 1983). Finally, the prototype model does not explain the effects of induction just discussed.

The list is quite similar for exemplar models. The major exemplar models (Kruschke 1992; Medin and Schaffer 1978; Nosofsky 1984) do not have any way by which features can be constructed or interpreted; the experiments testing these models have almost always used simple, artificial stimuli with only a few features. Knowledge has no part in the learning or categorization rules used by these models, so without some modification, they cannot account for the learning and categorization effects of knowledge reported above.

Exemplar models have not generally been extended to explain induction, so it is worth considering whether they might be able to account for the results of Kalish and Gelman (1992), Proffitt et al. (2000), and Ross and Murphy (1999), described earlier. Suppose that induction from one category member to another involves the retrieval of similar exemplars. For example, imagine that you are asked whether property X of cereal is also true for milk or noodles. One way to answer this question is to retrieve examples of cereal and of milk and see whether they are similar (i.e., have the same properties). To the degree that they are similar, then, one might respond positively to this induction question. Such a rule is parallel to the categorization rule used by the Context Model (Medin and Schaffer 1978) and its descendants, and it would seem to predict many of the data found in classic studies of induction (e.g., the typicality and similarity effects of Rips 1975—see Osherson et al. 1990). By itself, however, this rule would not explain the knowledge effects shown here, because it would not predict reversals of inductions from cereal to milk, compared to inductions from cereal to noodles, for different properties. The similarity of cereal to milk would determine the same induction strength for all properties, and so this could not change depending on the property as Ross and Murphy (1999) and Heit and Rubinstein (1994) found. Although the similarity rule could perhaps be modified to be context-sensitive, it is not clear how to make similarity depend on the induced property in just the right way, unless one incorporated a reasoning process of the sort the knowledge approach includes. The ecological and causal-reasoning processes used by Proffitt et al.'s experts are clearly beyond the scope of this exemplar comparison process. So, the traditional exemplar model does not immediately predict knowledge effects in induction.
It is clear, then, that the traditional models will have to be augmented or modified in order to account for all these results. Some progress is being made on this front. Heit (1994) suggested that exemplar models could incorporate knowledge during learning by representing knowledge as a number of previously encountered exemplars. For example, in his experiment, subjects learned about how many people in City W were shy and often attended parties. Heit proposed that the subjects already knew people who are shy or not shy, and who often or seldom attend parties. That is, they already had some knowledge of how often these features co-occurred. Under the assumption that exemplars with the expected feature pairs are more frequent than the unexpected ones (e.g., that shy people typically do not attend parties), Heit found that an exemplar model could account for his results. This proposal, then, has the advantage of using a standard psychological model to represent both empirical learning and knowledge effects, although it does address only one form of knowledge influence.

Using exemplars to represent prior knowledge could work in some cases but seems implausible in others. The above example used a pair of features whose relation is already known, but sometimes the knowledge accessed is abstract or a generalization that is inferred from specific examples. For example, I can understand that a flying squirrel flies (or glides, really) because of the folds of skin between its body and limbs, because this mechanism is analogous to that used by other flying and gliding entities. However, other exemplars like birds, planes, parachutes, and so on, do not have these folds or this exact gliding capability—it is only by analogy or through a generalization of those exemplars that I can understand this new category. Thus, simply having exemplars would not be sufficient to explain the flying squirrel—a more powerful inferential or abstraction mechanism would be needed. As a thought experiment, consider the feature pairs: shy-often talks in videoconferences and shy-often talks in online chatgroups. Are they equally likely? The first seems much less plausible than the second to me, because I expect shy people to be more forthcoming when they are not physically present or perceptible to others. However, I do not have direct experience with shy people in either situation, and so I must rely on general knowledge to draw that inference, rather than on known exemplars. As will be discussed below, most representations of knowledge in psychology and artificial intelligence assume that knowledge is about whole classes of entities rather than about individuals: Wings are useful for flying because of the properties of wings in general and the laws of aerodynamics in general, rather than because of the properties of a lot of exemplars. This point is not to deny that Heit's proposal could well account for a subset of knowledge effects, but how widely it can be applied is less clear.
Figure 6.4
A simplified depiction of Heit and Bott's (2000) Baywatch model of category learning. The input features represent properties of the objects, and Categories A and B are the to-be-learned categories. The prior knowledge (PK) nodes are the already known concepts of church and office buildings.

In a later paper, Heit and Bott (2000) developed a connectionist model of knowledge effects. They had taught their subjects concepts of buildings reminiscent of churches and office buildings. The features related to these familiar concepts were learned significantly faster than those that were unrelated to the themes. They constructed a 3-layer connectionist model of this task, in which the input layer was the features used in the items, the output layer was the two categories, and the hidden layer contained the prior knowledge (PK) nodes of church and office building. As shown in figure 6.4, all of the input features were directly connected to the output nodes, initially at very weak levels; in addition, the input features related to the prior knowledge were strongly connected to the PK nodes. For example, if an input feature was "has candles," it would be connected to the church building node. The categories, then, could be learned through two routes: direct associative learning between the input and category (output) nodes, and indirectly through the PK nodes. (Note that this system is a prototype model, as it represents the strength of relations between features and categories, rather than configurations of features that might define an exemplar.)
During learning, this model learned empirically by associating the input features to the two categories (e.g., “has candles” was associated to category B). It also learned that the PK nodes were related to the categories, that category B was like a church and category A was like an office building. As a result, the model learned the knowledge-related features better than the neutral features, as the subjects did. “Has candles” directly activated category B and indirectly activated it through the church building node, whereas a neutral feature like “near a bus stop” activated its category only through the direct route. One might complain that the PK nodes were somewhat rigged, as they just happened to correspond to the two categories to be learned. However, Heit and Bott showed that if multiple knowledge nodes were present, only some of which were relevant to the categories, the model did not become confused. Because only the relevant nodes were consistently associated to the categories, only their links to the categories were learned; the irrelevant prior concepts had no effect.

This sort of model is a good start for explaining knowledge effects. Although it is unlikely that PK nodes of exactly this sort are generally available (normally, the concepts being learned are not as similar to known concepts), one can easily imagine other kinds of generalizations and relations among features that might be represented in hidden nodes and that therefore could aid category learning. One possible problem with connectionist models of this sort, as Heit and Bott (2000) and Kaplan and Murphy (2000) both note, is that they predict that knowledge will reduce learning of features that are not related to it: Learning some features comes at the expense of others. However, this result has not been empirically found in human concept learning. Surprisingly, unrelated features are learned equally well by subjects who do and do not have prior knowledge (Kaplan and Murphy 2000). This result is difficult to obtain in connectionist models that use error-driven learning. (Rehder and Murphy 2001 have recently developed a model that expands Heit and Bott’s framework by using a more flexible representation of knowledge. It is able to learn knowledge-unrelated features while simultaneously showing the knowledge advantages found in experiments.)

Heit and Bott’s (2000) model has considerable promise, then. Ultimately, a complete model of the learning process will have to do more to represent the relevant knowledge, perhaps through addition of an inferential mechanism of some kind. Furthermore, this sort of model is focused specifically on the learning of known features and, like all current models, does not address issues of feature construction, knowledge in induction, and the like. However, in order to develop a model that does all these things, we must first have a successful model that does one or two of
them, and that is the goal for the field at present. Heit and Bott provide a useful discussion of various ways that knowledge could be incorporated into different kinds of models, and the interested reader should consult that article.

In short, we are just making a start at building theories and computational models that integrate empirical and knowledge-based learning. It might be useful to consider what a complete model would look like, so we can see the task ahead of us. Ignoring both practical and theoretical problems in developing such a system, how do the current data suggest that knowledge should be integrated into a learning model? First, an integrated model would have to represent a number of prior concepts at the beginning of the learning task, rather than modeling the learning of one or two concepts in a tabula rasa. Since similar concepts are a source of features and feature weights, they need to be present in order to give the new concepts a starting point. Learners also place special attention on properties that seem likely to be informative. So, in learning about a whingelow, the system must already know about other mammals in order to know which features to attend to and which generalizations to make from the observed examples. More difficult to implement, but probably also necessary would be a form of reasoning or inference engine. Somehow these models need to be able to use more general knowledge to draw inferences. These may not have to be very difficult or intelligent inferences, but, at the least, obvious ones such as “astronauts would live in a building in outer space, but divers would live in an underwater building.” (However, it should be pointed out that the history of Artificial Intelligence tells us that such inferences are notoriously difficult to draw, so working models would probably have to confine themselves to a simple domain with only a small number of facts to deal with.)

Why is this whole reasoning apparatus necessary? Some researchers have suggested that weights on dimensions at the start of learning could represent some prior knowledge (Kruschke 1993). For example, perhaps one could give greater attention to shape than color, since shape seems to be a more important dimension for category learning. However, this kind of solution will not be sufficient, as different features are important for different categories (see Kelemen and Bloom 1994). It cannot be a weight on the dimension itself, then, that accounts for this, because the weight would not change with different categories. (If you did not weight dimensions differently, you might pay a lot of attention to the present location of the whingelow, because present location is important for baseball positions. Then you would end up hypothesizing that whingleows are found only in the zoo, or only in San Diego, because that’s the only place you saw them.) Especially for very new kinds of categories (space vehicles or very deep sea creatures, for example), the sys-
tem could not rely on prior feature weights or exemplars but must be able to draw inferences from more general knowledge. Another problem with simply storing weights for each feature is that knowledge is also needed to tie together the features within a category. This can be seen in experiments that contrast the same features when they can be related together vs. when they cannot (Murphy and Alloppena 1994; Spalding and Murphy 1996; Wattenmaker et al. 1986). Such results obviously cannot be explained by feature weights, which are the same across conditions. However, it is also likely that frequently made inferences will be encoded into concepts so that they do not need to be drawn de novo every time.

General knowledge plus an inference engine could be used again during categorization. Knowledge would provide greater weight to some features of the item being categorized than to others, even if both were empirically found to be related to category membership. Furthermore, such knowledge could be used to infer unseen features (e.g., “that vehicle must have an engine, because it’s moving on its own”). I should note, however, that the use of knowledge during categorization of well-learned categories is perhaps the least well demonstrated of the knowledge effects described earlier. If knowledge influences the learning process so that the representation of the concept is affected, it is not so clear that one must also access this domain knowledge during categorization—it could already be incorporated into the concept (see discussion of categorization above). For example, after seeing the whingleow, you could have encoded quite a bit about its shape and behaviors, but not so much about the time of day that you saw it or the number of them in the cage. If you did so, then when deciding whether another animal is a whingleow, you would not need to consult domain knowledge to decide whether shape or time of day are critical, because this information would already be implicit in the concept. Thus, it is possible that a formal system could get by without using knowledge during the categorization process itself, if it is incorporated during learning. Of course, this consideration would not apply to ad hoc categories (Barsalou 1983), which are essentially made up on the spot. Here, categorization is clearly a knowledge-intensive process, but perhaps these cases are out of the range of category-learning models.

The Wisniewski and Medin (1994) results are probably the most difficult for the prototype and exemplar models to accommodate. The root assumption of these models is empiricist: They assume that exemplars present their features to the learner, which then are associated to a category. This observation leads to a representation of the category in terms of the observed properties. The Wisniewski and Medin study undermines this whole perspective, however, by suggesting that the features are not always present in the observed items but can instead be constructed,
Knowledge View

There will be little suspense in concluding that something called “the knowledge approach” is the winner in a chapter entitled “Knowledge Effects.” The phenomena reviewed in this chapter are exactly the ones that have caused some researchers to conclude that concept acquisition and use involves prior knowledge to a significant degree. It would be tedious to review all the phenomena and then point out that each one seems to require reference to domain knowledge. Therefore, in this section, I will consider more critically what the knowledge approach is and how it must handle the variety of effects discussed in this chapter.

As was mentioned in chapter 3, the knowledge approach is still somewhat incomplete. It has not been instantiated in a computational model to any significant degree, unlike the exemplar approach in particular. The statement that knowledge is heavily involved in many aspects of concept acquisition and use (as I have just reviewed) is now one that has extremely strong support. However, it takes much more than that statement to make a complete theory.

There are two reactions one could have to the kinds of results discussed in this chapter. The first, most aggressive approach would be to argue that the empirical theories are inadequate, since they do not account for the knowledge effects that have been demonstrated, and therefore they should be rejected. Instead, a new theory that relies more heavily on knowledge must be developed. This is certainly one possible conclusion, but at this stage, it seems premature for two reasons. The first is that the empirical phenomena discussed in other chapters, such as typicality, basic categories, and exemplar effects are extremely reliable, but they do not require prior knowledge. Clearly, people can learn categories in artificial and unfamiliar domains, and it is likely that at least some of the same learning mechanisms are involved in the learning of categories in knowledge-rich domains. Simply rejecting the more empirical approaches, then, provides no way of accounting for this commonality. Second, no knowledge-based approach has been proposed that could actually replace the empirical views. There is currently no general model based on domain knowledge, inference and reasoning, and so on, that can explain the sorts of results that the prototype and exemplar models have focused on. Given that, rejection of those models would be premature at this stage.

The second way to respond to the results presented in this chapter would be to try to fill in the gaps of the more empirical approaches. In the previous section, I argued that the learning models needed to be integrated with a knowledge base and reasoning process that used that knowledge during learning, categorization, and induction. The knowledge approach, then, could well be assigned the task of providing that
in part as a function of prior knowledge. Similarly, the work of Schyns and colleagues (Schyns et al. 1998; Schyns and Murphy 1994; Schyns and Rodet 1997) has shown that the category-learning process itself helps to define what the features are. This makes it very difficult for both the exemplar and prototype views, because the features that are being associated to the category are a moving target. You can't count how often each one is associated to each category (or see which exemplars have which features) if the features themselves are changing from trial to trial.

How could this sort of effect be incorporated into the formal learning models? If one allows only a pre-set list of features such as “green,” “square,” “large,” and so on, then there is no possibility of modifying them or adding to the set. In some sense, what these effects require is the entire perceptual system to be an input to the learning module. In Schyns's experiments, for example, the stimuli are rather ambiguous entities that could be perceptually divided in a number of different ways. The learning system needs to have access to the multiple representations constructed during perception, so that it can select among the possible ones in order to associate the correct one to the category. If one section of the blob keeps coming up in one category but not the other, this has to be identifiable as a coherent section in the perceptual system and then associated to the correct category by the learning component. (See Schyns 1991 for a related computational model.)

Even more complex are Wisniowski and Medin's findings with children's drawings, which reveal an interaction between the perceptual input and prior expectations of the category. The perceptual system must represent such drawings in a flexible enough manner that dots on the front of a shirt could be perceived as buttons by one subject and as a tie by another—the decision being made by knowledge that prompted the subjects to look for detail or city clothing. This system would be extremely complex, though, because not only does the knowledge influence the perceptual identification of properties, learning influences whether these encodings are maintained. That is, after identifying something as a detailed drawing, when a subject learns that it was not made by a creative child, he or she might decide that the drawing was not in fact detailed enough. So, there is a three-way interaction of the input, the prior expectation, and the feedback.

Unfortunately, very little of this fits in with current formal modeling. For proponents of prototype and exemplar theory, it can only be hoped that when such models are developed, the learning mechanisms and representations that they have proposed still have a place at the center of the learning process. But it is premature to say whether that will be the case or not, given that the field seems far from incorporating all these effects in any kind of model.
1999; Wattenmaker 1995; Wisniewski 1995). If the critique of those traditional models of category learning that ignore knowledge is to go forward, however, it will be necessary for proponents of this view to provide similarly detailed models of their own. By the same token, proponents of formal theories need to develop their own models to provide an account of all the phenomena described here. This is in fact how the most recent proposals of Heit (1994; Heit and Bott 2000) developed.

The question that has yet to be answered is exactly how knowledge changes the learning process and the resulting concept representation. Early proposals (Murphy and Alloppena 1994) suggested that the knowledge simply became the category representation to some degree. For example, once you figured out “Category 2 is arctic vehicles” (see table 6.1), you didn’t then devote attention to learning the category’s properties or exemplars, but simply represented the category as arctic vehicles. On this account, then, knowledge pre-empted empirical learning. But later findings show that this proposal is clearly wrong—knowledge doesn’t hurt the learning of statistical properties of the stimuli; in some cases, it may even help it (see especially Kaplan and Murphy 1999, 2000; Spalding and Murphy 1999). For example, subjects learn the frequency of features related to knowledge somewhat better than subjects do who do not have knowledge (Spalding and Murphy 1999). Theories of concept learning will need to be developed that can explain how knowledge simultaneously (1) greatly speeds category learning, (2) greatly aids learning of features related to the knowledge, (3) yet does not impair statistical learning of those features (or benefits it), and (4) does not hurt learning of features unrelated to knowledge. This is a surprising pattern of results, which is a challenge to all current theories.

Models from Outside Psychology

One place to look for inspiration on how to incorporate knowledge is the literature on machine learning in artificial intelligence (AI). More traditional learning algorithms have been called Similarity-Based Learning (SBL), because they learn categories based on the sharing of properties among items in the same class. These are analogous to the traditional prototype and exemplar models in psychology. SBL is contrasted with Explanation-Based Learning (EBL), which is related to the knowledge approach in psychology. In early EBL systems (e.g., DeJong and Mooney 1986), the computer started out with a bunch of knowledge, which could be thought of as premises. The system was exposed to a single category member, along with some kind of description of what the category was supposed to be (e.g., a tool to accomplish some function; a treatment for a particular disease). The system then used its
component of a complete learning model; researchers in this approach should be saying what kind of knowledge people use, how it influences learning, how people use it during categorization, and so on. There are many gaps here that need filling. For example, we do not have anything like a taxonomy of different types of background knowledge and an understanding of how each type might influence learning (see Murphy 2000, and Wilson and Keil 2000, for discussion). And different research projects have drawn on different knowledge sources. For example, Wattenmaker (1995) activated knowledge structures by telling subjects that the items resembled categories that they already knew (see also Ahn 1990; Heit and Bott 2000; Wattenmaker et al. 1986, Experiment 6). In contrast, Pazzani (1991) evoked knowledge by using a pre-existing causal schema (see also Ahn 1998; Ahn et al. 2000; Rehder and Hastie 2001). Other studies have used categories that were “thematic related,” in that the properties could all be related to a similar function or environment (e.g., Murphy and Allopenna 1994; Spalding and Murphy 1999). However, not all of these relations were clearly causal (e.g., arctic vehicles being white) or any other single type of relation. Although all these studies show the use of “knowledge,” it is not at all clear that the knowledge involved is the same kind of thing, or that it is influencing learning in the same way. In short, this is a promising avenue for exploration.

A similar point could be made about categorization. I discussed above two different ways that knowledge might affect categorizations: (1) indirectly, by changing the concept representation during learning (i.e., before categorization); or (2) directly, by activating the knowledge during the category judgment. It is likely that both effects occur at least some of the time. For example, for fast visual categorization (as in Lin and Murphy 1997; Palmeri and Blalock 2000), the indirect influence seems most likely, but for ad hoc categories (Barsalou 1983) and the medical-experiment-gone-horribly-wrong categorization problems used by Keil (1989) and Rips (1989), it seems likely that the knowledge is used during the decision process itself. We now must ask what distinguishes these cases, and what the limits are of each kind of knowledge use.

Much of the research within the knowledge approach has been devoted toward showing that knowledge does indeed have an effect and that purely formal models of categorization are not sufficient. This point has been made very forcefully, and now it is necessary to go beyond it in providing explicit, detailed accounts of how knowledge is involved in these processes. More recent research has in fact been designed to differentiate hypotheses about how knowledge influences concept use (e.g., Heit 1994, 1998; Kaplan and Murphy 1999, 2000; Spalding and Murphy
knowledge to attempt to prove that the object was in fact in the category. The program could also generalize the proof so as to specify the range of observable properties that would satisfy it. This proof then became the representation of the category as a whole. In this way, the system would not have to go through a new proof with every new category member that was encountered, and it would not be misled by idiosyncratic properties of the learning exemplar.

This description may be a bit puzzling to psychologists, because it seems to assume that the system already knows what the category is, or else it could not have derived a proof, especially after observing only one exemplar. The system knows about the entire domain and also knows something about the function or goal of the category itself. (In Ahn, Brewer, and Mooney 1992, for example, subjects were provided with the purpose of the category before being exposed to the exemplar. In DeJong and Mooney 1986, the learning program already had a set of knowledge about human behavior and then had to learn a category from a narrative that exemplified a particular category of human behavior.) This situation still requires further learning, because even if one has general knowledge of a category, such as its function, one must still learn which particular properties are characteristic of it and be able to identify members based on perceptual properties. However, I think that it is fair enough to say that EBL models of this sort do not capture the typical situations in which one is confronted with objects and their category labels, or in which one has observed a number of objects and then realizes that they form a category. In most such cases, the learner does not already have a function available for the object. Also, learners can and do learn unexpected properties of objects as a result of induction from multiple examples.

More recent models, however, have focused on situations that are somewhat more realistic, or, at least, somewhat more like psychology experiments. For example, Pazzani (1991) developed a computer program called Post-Hoc that simulated the performance in his experiment (described above). Post-Hoc developed hypotheses about the correct category representation by starting from its knowledge and then looking for relevant properties in the examples. For example, if the program was trying to learn the category of balloons that would inflate, it would then look for the properties of being stretched and being inflated by an adult. If a positive example had one of those features, it would be incorporated into the hypothesis about the category. It also had a fairly simple learning mechanism that would fix hypotheses that were found to be incorrect based on later evidence. Pazzani presented evidence that his model did account for important aspects of the human data.

Post-Hoc does have some interesting ideas for how knowledge-based learning may work, but it is incomplete in a number of respects. First, it only learns classical
categories in which features are related by conjunction or disjunction. It would find it very difficult to learn many of the family-resemblance categories used in other experiments. Second, the model makes the strong claim that knowledge features are learned to the exclusion of other features (when the knowledge is correctly related to category membership). However, more recent evidence shows that people learn other features as well. For example, Kaplan and Murphy (2000) found that providing subjects with knowledge did make them focus more on the knowledge-related features. However, their learning of other features was exactly the same as that of subjects in the neutral group, and was fairly good overall. Wisniewski (1995) and Heit and Bott (2000) have also demonstrated learning of features unrelated to category knowledge. As noted earlier, this result is a puzzling one for many proposals of how knowledge influences learning, because it requires that knowledge speed learning and help some features but that the other features still be learned. This seems to be a realistic aspect of concept learning, since people learn not only knowledge-related features of everyday categories (e.g., birds have wings in order to fly), but also features that are not clearly related to that knowledge (e.g., birds have beaks).

IOU was designed by one of the originators of EBL approaches to learning, Raymond Mooney, to address some of these limitations of a purely knowledge-based approach. Mooney (1993) pointed out that prior theory can typically only explain some proportion of a concept’s features, and that some other mechanism will be necessary to discover its other features. In the IOU (Induction Over the Unexplained) system, domain knowledge is first used to explain as much of the observed examples as possible. However, if this explanation does not result in correct categorization, it is augmented by an empirical learning system that learns additional features.

For example, suppose the system were trying to learn about cups, and it had the idea that cups are drinking vessels, but didn’t know anything else about them (Mooney 1993). It would first attempt to explain as many features as possible based on its knowledge and the drinking vessel function (as in the usual EBL system). Now, some features of observed cups could be easily explained (cylindrical shape), but others might not be so easily explained, by the system, at least (their typical width and height). Furthermore, the explanation of a drinking vessel might pick out not only cups but also bowls and glasses. Therefore, the system must be exposed to cups and contrast categories, labeled with their category membership, in order to learn to separate cups from related items on an empirical basis. The explainable features are ignored in the empirical process, and only the unexplained ones are submitted to it. In one version of the program, the empirical learning component notices all the (unexplained) features that are common to the observed examples.
Mooney (1993) reviews how the model can explain aspects of psychological experiments, such as those of Wisniewski (1995) and Ahn et al. (1992). Indeed, the model is an important step in unifying empirical and knowledge-based learning. Like other EBL systems, it has the limitation that one must already have some description of the category to allow one to explain the observed features. It is not clear if the model will avoid the problems I raised for the Post-Hoc model. For example, it also seems to require that the explainable features be present in all items. IOU can certainly explain giving more weight to knowledge-related features, which are learned first; and it also accommodates the learning of purely empirical features. The model is not designed to accomplish unsupervised learning, but it does not seem impossible to adapt it to such a situation. Thus, this model has some promise, if it can overcome the issue of having to know the goal or other information about the category in advance. Although it does not address the more difficult issues of feature construction and induction, that is beyond any model at this time.

AI models hold promise for helping us understand the interaction of knowledge and empirical learning processes. Their main shortcomings as psychological accounts are twofold. First, they are sometimes not tested on psychological data but rather on in-principle problems, to see if they form reasonable categories. (This is less true with the particular models I have been reviewing in this chapter, which have addressed at least some psychological data.) Second, the models often have some assumptions that are psychologically implausible, for example, a reliance on predicate logic and classical rules in the knowledge component. These AI proposals are a good jumping-off point, but more specifically psychological theories of the interaction of knowledge and empirical learning still need to be developed.

Perhaps it is worth pointing out that these AI approaches all appear to be prototype models. That is, they produce descriptions that apply to the category as a whole, rather than learning individual exemplars. This is probably based on the belief that knowledge tends to be about whole categories rather than individual entities, and it is most convenient to represent it that way. In any case, these models therefore carry whatever positive or negative baggage that prototype models have in general.

**Future Directions**

I have been emphasizing the development of more complete models in this discussion, but there is still much psychological research to be done in order to increase our understanding of how knowledge effects work, and how they are integrated with empirical learning. At the present moment, the most critical question to my
mind is the issue of how the learning process itself is changed by prior knowledge and how that in turn leads to changes in the category representation (see discussion at the end of the "Knowledge View" section above). The knowledge and statistical learning are apparently interacting in some way, but it is not obvious how.

As remarked in a number of places, there also needs to be further work done on knowledge in categorization, as the number of studies on this is somewhat slim, and important theoretical alternatives need to be distinguished. And finally, the influence of knowledge on feature construction is a fascinating question that has been little studied since Wisniewski and Medin's (1994) groundbreaking work.

In sum, there is still much to discover beyond the initial "does knowledge make a difference?" questions that motivated the earliest studies on this topic.