overturn it. It was only after such data were well established that classical theorists proposed reasons for why these results might not pose problems for an updated classical view.

As I mentioned earlier, there is no specific theory of concept representation that is based on the classical view at the time of this writing, even though there are a number of writers who profess to believe in this view. The most popular theories of concepts are based on prototype or exemplar theories that are strongly unclassical. Until there is a more concrete proposal that is "classical" and that can positively explain a wide variety of evidence of typicality effects (rather than simply criticize the arguments against it), we must conclude that this theory is not a contender. Thus, although it pops up again from time to time, I will not be evaluating it in detail in the remainder of this book.

3 Theories

As described in the previous chapter, the classical view has taken a big fall. Into this vacuum other theories developed that did not assume that concepts were represented by definitions and so were not subject to the problems the classical view suffered. This chapter will consider the three main kinds of theories that arose after the downfall of the classical view. The goal here is not to comprehensively evaluate these theories, as that can be done only over the course of the entire book, after the complete range of relevant data has been presented. Instead, this chapter will introduce the three general approaches that are most current in the field and explain how they deal with the typicality phenomena that caused such a problem for the classical view.

The Prototype View

One of the main critics of the classical view of concepts was Eleanor Rosch, who provided much of the crucial evidence that revealed the shortcomings of a definitional approach to concepts. Rosch's writings also provided the basis for a number of the early alternatives to the classical view, all under the rubric of the *prototype view*.

A number of readers interpreted Rosch as suggesting that every category is represented by a single prototype or best example. That is, perhaps your category of dogs is represented by a single ideal dog, which best embodies all the attributes normally found in dogs. I gave such an interpretation in the previous chapter as one way of understanding the existence of typicality. For example, very typical items would be those that are similar to this prototype; borderline items would be only somewhat similar to this prototype and somewhat similar to other prototypes as well. (The dot-pattern experiments of Posner and Keele 1968, 1970, encouraged this interpretation as well, as their categories were constructed from a literal prototype.)

Rosch, however, explicitly denied that this was her proposal (Rosch and Mervis 1975, p. 575), though it must be said that her writing often encouraged this interpretation. She preferred to be open-minded about how exactly typicality structure was represented, focusing instead on documenting that it existed and influenced category learning and judgments in important ways.

The idea that a single prototype could represent a whole category is questionable. For example, is there really an "ideal bird" that could represent all birds, large and small; white, blue, and spotted; flightless and flying; singing, cackling, and silent; carnivorous and herbivorous? What single item could pick out penguins, ostriches, pelicans, hummingbirds, turkeys, parrots, and sparrows? It seems unlikely that a single representation could encompass all of these different possibilities (Medin and Schwanenflugel 1981). Furthermore, a single prototype would give no information about the variability of a category. Perhaps some categories are fairly narrow, allowing only a small amount of variation, whereas others have very large variation (e.g., compare the incredible variety of dogs to the much smaller diversity of cats). If each category were represented by a single "best example," there would be no way to represent this difference (see Posner and Keele 1968, for an experimental demonstration).

In short, the notion of a single prototype as a category representation, which I'll call the *best example* idea, has not been very widely adopted. Instead, the prototype view proposed by Rosch has most often been interpreted as a summary representation that is a description of the category as a whole, rather than describing a single, ideal member. The view I'll discuss was proposed by Hampton (1979) and was fleshed out by Smith and Medin (1981), although its roots lie in Rosch and Mervis (1975).

A critical component of the prototype view is that it is a *summary representation* (just why this is critical will be apparent when the exemplar view is discussed): The entire category is represented by a unified representation rather than separate representations for each member or for different classes of members. This may sound just like the single best example idea that I have just criticized, but you will see that this representation is considerably more complex than that.

The representation itself could be described in terms much like Rosch and Mervis's (1975) family-resemblance view. The concept is represented as features that are usually found in the category members, but some features are more important than others. It is important for weapons that they be able to hurt you, but not so important that they be made of metal, even though many weapons are. Thus, the feature "can do harm" would be highly weighted in the representation, whereas the feature "made of metal" would not be. Where do these weights come from? One possibility is that they are the family-resemblance scores that Rosch and Mervis derived (see previous chapter). That is, the more often a feature appears in the category and does not appear in other categories, the higher its weight will be. Unlike a best-example representation, this list of features can include contradictory features with their weights. For example, the single best example of a dog might be short-haired. However, people also realize that some dogs have very long hair and a few have short hair. These cannot all be represented in a single best example. The feature list would represent this information, however. It might include "short hair," and give it a high weight; "long hair," with a lower weight; and "hairless" with a very low weight. In this way, the variability in a category is implicitly represented. Dimensions that have low variability might have a single feature with high weight (e.g., "has two ears" might have a high weight, and "has three ears" would presumably not be listed at all). Dimensions with high variability (like the colors of dogs) would have many features listed ("white," "brown," "black," "orange-ish," "spotted"), and each one would have a low weight. Such a pattern would implicitly represent the fact that dogs are rather diverse in their coloring. This system, then, gives much more information than a single best example would.

One aspect of this proposal that is not clearly settled yet is what to do with continuous dimensions that do not have set feature values. So, how does one represent the size of birds, for example: as "small," "medium," and "large," or as some continuous measurement of size? If it is a continuous measurement, then feature counting must be somewhat more sophisticated, since items with tiny differences in size should presumably count as having the same size feature, even if they are not identical. Perhaps for such continuous dimensions, what is remembered is the average rather than the exact features. Another idea is that features that are distinctive are counted, whereas those that are close together are averaged. So, for categories like robins, the size differences are small enough that they are not represented, and we only remember the average size for a robin; but for categories like birds as a whole, we do not average the size of turkeys, hawks, robins, and wrens, which are too diverse. There is some evidence for this notion in category-learning experiments (e.g., Strauss 1979), but it must be said that a detailed model for how to treat such features within prototype theory seems to be lacking.

If this feature list is the concept representation, then how does one categorize new items? Essentially, one calculates the similarity of the item to the feature list. For every feature the item has in common with the representation, it gets "credit" for the feature's weight. When it lacks a feature that is in the representation, or has a feature that is not in the representation, it loses credit for that feature (see Smith and Osherson 1984; Tversky 1977). After going through the object's features, one adds up all the weights of the present features and subtracts all the weights of its features that are not part of the category.[1] If that number is above some critical value, the *categorization criterion*, the item is judged to be in the category; if not, it is not. Thus, it is important to have the highest weighted features of a category in order to be categorized. For example, an animal that eats meat, wears a collar, and is a pet might possibly be a dog, because these are all features associated with dogs, though not the most highly weighted features. If this creature does not have the shape or head of a dog, does not bark, does not drool, and does not have other highly weighted dog features, one would not categorize it as a dog, even though it wears a collar and eats meat. So, the more highly weighted features an item has, the more likely it is to be identified as a category member.

This view explains the failure of the classical view. First, no particular feature is required to be present in order to categorize the item. The inability to find such defining features does not embarrass prototype theory the way it did the classical view. So long as an item has enough dog features, it can be called a dog—no particular feature is defining. Second, it is perfectly understandable why some items might be borderline cases, about which people disagree. If an item has about equal similarity to two categories (as tomatoes do to fruit and vegetable), then people may well be uncertain and change their mind about it. Or even if the item is only similar to one category, if it is not very similar—in other words, right near the categorization criterion—people will not be sure about it. They may change their mind about it on a different occasion if they think of slightly different features or if there's a small change in a feature's weight. Third, it is understandable that any typical item will be faster to categorize than atypical items. Typical items will have the most highly weighted features (see Barsalou 1985; Rosch and Mervis 1975), and so they will more quickly reach the categorization criterion. If you see a picture of a German shepherd, its face, shape, size, and hair all immediately match highly weighted values of the dog concept, which allow speedy categorization. If you see a picture of a sheepdog, the face, length of hair, and shape are not very typical, and so you may have to consider more features in order to accumulate enough weights to decide that it is a dog.

Recall that Hampton (1982) demonstrated that category membership judgments could be intransitive. For example, people believe that Big Ben is a clock, and they believe that clocks are furniture, but they deny that Big Ben is furniture. How can this happen? On the prototype view, this comes about because the basis of similarity changes from one judgment to the other. Big Ben is a clock by virtue of telling time; clocks are furniture by virtue of being objects that one puts in the home for decoration and utility (not by virtue of telling time, because watches are not considered furniture). However, Big Ben is not similar to the furniture concept, because it isn't in the home and is far bigger than any furniture. Thus, concept *A* can be similar to concept *B*, and *B* can be similar to *C*, and yet *A* may not be very similar to *C*. This can happen when the features that *A* and *B* share are not the same as the features that *B* and *C* share (see Tversky 1977). On the classical view, this kind of intransitivity is not possible, because any category would have to include all of its superset's definition, and so there is no way that deciding that something is a clock would not also include deciding that it was furniture.

Smith and Medin (1981) discuss other results that can be explained by this feature-listing model. Most prominent among them are the effects of false relatedness: It is more difficult to say "no" to the question "Is a dog a cat?" than to "Is a dog a mountain?" I will leave it as an exercise for the reader to derive this result from the feature list view.

More Recent Developments

Unlike the other views to be discussed in this chapter, the prototype view has not been undergoing much theoretical development. In fact, many statements about prototypes in the literature are somewhat vague, making it unclear exactly what the writer is referring to—a single best example? a feature list? if a feature list, determined how? This lack of specificity in much writing about prototype theory has allowed its critics to make up their own prototype models to some degree. As we shall see in chapter 4, many theorists assume that the prototype is the single best example, rather than a list of features, even though these models have very different properties, for real-life categories, at least.

Feature combinations. The view taken by Rosch and Mervis (1975), Smith and Medin (1981), and Hampton (1979) was that the category representation should keep track of how often features occurred in category members. For example, people would be expected to know that "fur" is a frequent property of bears, "white" is a less frequent property, "has claws" is very frequent, "eats garbage" of only moderate frequency, and so on. A more elaborate proposal is that people keep track not just of individual features but configurations of two or more features. For example, perhaps people notice how often bears have claws AND eat garbage, or have fur AND are white—that is, combinations of two features. And if we propose this, we might as well also propose that people notice combinations of three features such as having claws AND eating garbage AND being white. So, if you saw a bear with brown fur eating campers' garbage in a national park, you would update your category information about bears by adding 1 to the frequency count for features "brown," "has fur," "eats garbage," "brown and has fur," "brown and eats garbage," "has fur and eats garbage," and "brown and has fur and eats garbage." This proposal was first made by Hayes-Roth and Hayes-Roth (1977) and was made part of a mathematical model (the *configural cue model*) by Gluck and Bower (1988a).

One problem with such a proposal is that it immediately raises the question of a computational explosion. If you know 25 things about bears, say (which by no means is an overestimate), then there would be 300 pairs of features to be encoded. (In general, for N features, the number of pairs would be $N * (N - 1)/2$.) Furthermore, there would be 2,300 triplets of features to encode as well and 12,650 quadruplets. Even if you stopped at triplets of features, you have now kept track of not just 25 properties, but 2,635. For any category that you were extremely familiar with, you might know many more features. So, if you are a bird watcher and know 1,000 properties of birds (this would include shapes, sizes, habitats, behaviors, colors, and patterns), you would also know 499,500 pairs of features and 166,167,000 feature triplets. This is the explosion in "combinatorial explosion." Not only would this take up a considerable amount of memory, it would also require much more processing effort when using the category, since every time you viewed a new category member, you would have to update as many of the pairs, triplets, and quadruplets that were observed. And when making a category decision, you couldn't just consult the 1,000 bird properties that you know about—you would also have to consult the relevant feature pairs, triplets, and so on.

For these reasons, this proposal has not been particularly popular in the field at large. Models that encode feature combinations have been able to explain some data in psychology experiments, but this may in part be due to the fact that there are typically only four or five features in these experiments (Gluck and Bower 1988a, pp. 187–188, limited themselves to cases of no more than three features), and so it is conceivable that subjects could be learning the feature pairs and triplets in these cases. However, when the Gluck and Bower model has been compared systematically with other mathematically specified theories, it has not generally done as well as the others (especially exemplar models), as discussed by Kruschke (1992) and Nosofsky (1992). In short, this way of expanding prototype theory has not caught on more generally. The question of whether people actually do notice certain *pairs* of correlated features is discussed at greater length in chapter 5.

Schemata. One development that is tied to the prototype view is the use of *schemata* (the plural of *schema*) to represent concepts. This form of representation has been taken as an improvement on the feature list idea by a number of concepts researchers (e.g., Cohen and Murphy 1984; Smith and Osherson 1984). To understand why, consider the feature list view described above. In this view, the features are simply an unstructured list, with associated weights. For example, the concept of bird might have a list of features such as wings, beak, flies, gray, eats bugs, migrates in winter, eats seeds, blue, walks, and so on, each with a weight. One concern about such a list is that it does not represent any of the relations between the features. For example, the features describing a bird's color are all related: They are different values on the same dimension. Similarly, the features related to what birds eat are all closely connected. In some cases, these features are mutually exclusive. For example, if a bird has a black head, it presumably does not also have a green head and a blue head and a red head. If a bird has two eyes, it does not have just one eye. In contrast, other features do not seem to be so related. If a bird eats seeds, this does not place any restriction on how many eyes it has or what color its head is, and vice versa.

A schema is a structured representation that divides up the properties of an item into dimensions (usually called *slots*) and values on those dimensions (*fillers* of the slots). (For the original proposal for schemata, see Rumelhart and Ortony 1977. For a general discussion of schemata see A. Markman 1999.) The slots have restrictions on them that say what kinds of fillers they can have. For example, the head-color slot of a bird can only be filled by colors; it can't be filled by sizes or locations, because these would not specify the color of the bird's head. Furthermore, the slot may place constraints on the specific value allowed for that concept. For example, a bird could have two, one or no eyes (presumably through some accident), but could not have more than two eyes. The slot for number of eyes would include this restriction. The fillers of the slot are understood to be competitors. For example, if the head color of birds included colors such as blue, black, and red, this would indicate that the head would be blue OR black OR red. (If the head could be a complex pattern or mixture of colors, that would have to be a separate feature.) Finally, the slots themselves may be connected by relations that restrict their values. For example, if a bird does not fly, then it does not migrate south in winter. This could be represented as a connection between the locomotion slot (which indicates how the bird moves itself around) and the slot that includes the information on migration.

Why do we need all this extra apparatus of a schema? Why not just stick with the simpler feature list? The answer cannot be fully given here, because some of the evidence for schemata comes from the topics of other chapters (most notably chapter

12). However, part of the reason is that the unstructured nature of the feature list could lead to some peculiar concepts or objects being learned. Any features can be added to the list, and there are no restrictions on one feature's weights depending on what other features are. So, if I told you that birds are generally carnivorous, and you read somewhere that birds were generally vegetarian, you could simply list the features in your concept, say, carnivorous, weight = .80; and vegetarian, weight = .80. The fact that these two features both having high weights is contradictory would not prevent a feature list from representing them. Similarly, if I represented both the features "flies" and "doesn't fly" for birds (since some do and some don't), there is nothing to stop me from thinking that specific birds have both these features. The intuition behind schema theory is that people structure the information they learn, which makes it easier to find relevant information and prevents them from forming incoherent concepts of this sort (for evidence from category learning in particular, see Kaplan 1999; Lassaline and Murphy 1998).

Another argument often made about feature lists is that they do not have the kinds of relations that you need to understand an entire object. For example, a pile of bird features does not make a bird—the parts need to be tied together in just the right way. The eyes of a bird are *above* the beak, placed *symmetrically* on the head, *below* the crest. This kind of information is critical to making up a real bird, but it usually does not appear in feature lists, at least as produced by subjects in experiments. Subjects may list "has eyes," but they will not provide much relational information about how the eyes fit with the other properties of birds. Nonetheless, people clearly learn this information, and if you were to see a bird with the beak and eyes on opposite sides of its head, you would be extremely surprised. Schemata can include this information by providing detailed relations among the slots.

In short, a feature list is a good shorthand for noting what people know about a category, but it is only a shorthand, and a schema can provide a much more complete picture of what people know about a concept. For some purposes, this extra information is not very relevant, and so it may be easier simply to talk about features, and indeed, I will do so for just that reason. Furthermore, in some experiments, the concepts have been constructed basically as feature lists, without the additional relational information that a schema would include. In such cases, talking about feature lists is sufficient. However, we should not make the mistake of believing too much in the concepts we make up for experiments, which probably drastically underestimate the complexity and richness of real-world concepts. Schemata may be a better description of the latter, even if they are not required for the former.

The Exemplar View

The theory of concepts first proposed by Medin and Schaffer (1978) is in many respects radically different from prior theories of concepts. In the exemplar view, the idea that people have a representation that somehow encompasses an entire concept is rejected. That is, one's concept of dogs is not a definition that includes all dogs, nor is it a list of features that are found to greater or lesser degrees in dogs. Instead, a person's concept of dogs is the set of dogs that the person remembers. In some sense, there is no real concept (as normally conceived of), because there is no summary representation that stands for all dogs. However, as we shall see, this view can account for behaviors that in the past have been explained by summary representations.

To explain a bit more, your concept of dogs might be a set of a few hundred dog memories that you have. Some memories might be more salient than others, and some might be incomplete and fuzzy due to forgetting. Nonetheless, these are what you consult when you make decisions about dogs in general. Suppose you see a new animal walking around your yard. How would you decide that it is a dog, according to this view? This animal bears a certain similarity to other things you have seen in the past. It might be quite similar to one or two objects that you know about, fairly similar to a few dozen things, and mildly similar to a hundred things. Basically, what you do is (very quickly) consult your memory to see which things it is most similar to. If, roughly speaking, most of the things it is similar to are dogs, then you'll conclude that it is a dog. So, if I see an Irish terrier poking about my garden, this will remind me of other Irish terriers I have seen, which I know are dogs. I would conclude that this is therefore also a dog.

As in the prototype view, there must also be a place for similarity in this theory. The Irish terrier in my yard is extremely similar to some dogs that I have seen, is moderately similar to other dogs, but is mildly similar to long-haired ponies and burros as well. It has the same general shape and size as a goat, though lacking the horns or beard. It is in some respects reminiscent of some wolves in my memory as well. How do I make sense of all these possible categorizations: a bunch of dogs, a few goats, wolves, and the occasional pony or burro? Medin and Schaffer (1978) argued that you should weight these items in your memory by how similar they are to the item. The Irish terrier is extremely similar to some of my remembered dogs, is moderately similar to the wolves, is only slightly similar to the goat, and only barely similar to the ponies and burro. Therefore, when you add up all the similarities, there is considerably more evidence for the object's being a dog than for its being anything else. (I will describe this process in more detail later.) So, it is not just the number of exemplars that an item reminds you of that determines how you categorize it; just as important is how similar the object is to each memory.

How does this view explain the phenomena that the prototype view explained? First, this theory does not say anything about defining characteristics, so the problems for the classical view are not problems for it. Second, the view has a natural explanation for typicality phenomena. The most typical items are the ones that are highly similar to many category members. So, a German shepherd is extremely similar to many dogs and is not so similar to other animals. A dachshund is not as similar to other dogs, and it bears a certain resemblance to weasels and ferrets, which count against it as a dog. A chihuahua is even less similar to most dogs, and it is somewhat similar to rats and guinea pigs, and so it is even less typical. Basically, the more similar an item is to remembered dogs, and the less similar it is to remembered nondogs, the more typical it will be. Borderline cases are items that are almost equally similar to remembered category members and noncategory members. So, a tomato is similar to some fruit in terms of its having seeds, being round with edible skin, and so forth, but is similar to some vegetables in terms of its taste and how it is normally prepared.

Typical items would be categorized faster than atypical ones, because they are very similar to a large number of category members, and so it is very easy to find evidence for their being members. When you see a German shepherd, you can very quickly think of many dogs you have seen that are similar to it; when you see a chihuahua, there are fewer dogs that are similar to it. Thus, the positive evidence builds up more quickly when an item is typical (Lamberts 1995; Nosofsky and Palmeri 1997). The case of category intransitivity is explained in a way similar to that of the prototype view. For example, Big Ben is similar to examples of clocks you have seen in many respects. Clocks are similar to furniture exemplars in different respects. But Big Ben is not very similar to most furniture exemplars (beds, dressers, couches, etc.), and so it does not reach the categorization criterion. Whenever the basis for similarity changes, the exemplar model can explain this kind of intransitivity.

In short, the exemplar view can explain a number of the major results that led to the downfall of the classical view. For some people, this view is very counterintuitive. For example, many people don't consciously experience recalling exemplars of dogs in deciding whether something is a dog. However, conscious experience of this sort is not in general a reliable guide to cognitive processing. Indeed, one typically does not have conscious experience of a definition or a list of features, either. Access to the concept representation is often very fast and automatic. Second, some people point out that they feel that they know things about dogs, *in general*, not just about

individual exemplars. This is a concern that will come up later as well. However, note that what you know about dogs in general may be precisely what is most common in the remembered exemplars. So, when you think of such exemplars, these general characteristics are the ones that come to mind. Finally, when you first learn a category, exemplar information may be all you encode. For example, if you went to the zoo and saw a llama for the first time, all you know about llamas would be dependent on that one exemplar. There would be no difference between your memory for the whole category and your memory for that one exemplar. If you saw another llama a few months later, you could now form a generalization about llamas as a whole (though the exemplar view is saying that you do not do this). But you would clearly also remember the two llama examples as separate items. When you saw the third llama, you might still remember parts of the first two llamas, and so on. The question is, then, when you have seen a few dozen llamas, are you forming a general description of llamas—as the prototype view says—or are you just getting a better idea of what llamas are like because you have more memories to draw on—as the exemplar view says? But at the very initial stages of learning, it seems that any theory would have to agree that you remember the individual exemplars, or else you would have no basis on which to form generalizations (see Ross, Perkins, and Tenpenny 1990, for a clever twist on this idea).

A final point is that the exemplar model requires that you have specifically categorized these memories. You can easily recognize a German shepherd as a dog because it is similar to other things you have identified as dogs. If you had just seen a lot of other German shepherds without knowing what they were, they couldn't help you classify this similar object. This requirement of explicit encoding will come up again later.

Similarity calculation. Medin and Schaffer (1978) were the first to propose an elaborate exemplar model of concepts. In addition to the exemplar aspect itself, they also introduced a number of other ideas that were taken up by the field. One aspect was the nature of the similarity computation used to identify similar exemplars. In the past, most researchers had considered similarity to be an additive function of matching and mismatching features (Tversky 1977). Medin and Schaffer, however, proposed a *multiplicative* rule, which had a number of important effects on how their model operated.

The first part of this rule (as for additive versions) requires us to identify which aspects of two items are shared and which are different. For example, consider Ronald Reagan and Bill Clinton. They share many aspects: Both are men, are

Americans, were born in the twentieth century, were Presidents of the United States, attended college, are married, and so on. They also differ in a number of aspects: Where they were born, their ages, their political philosophies, their presidential actions, whether they have been divorced, what kinds of clothes they wear, and so on. For each of these matching and mismatching features, we need to decide two things. First, how important is this dimension to determining similarity? So, how important is it for two people to have the same or different sex? How important is age or marital status? How important is political philosophy? We need to decide this so that trivial differences between the two men don't swamp our decision. For example, perhaps neither man has been to Nogales, Arizona, but this should not count as a very important similarity. Perhaps Ronald Reagan has never played bridge, but Bill Clinton has. This does not seem to be an important difference. Second, for the mismatching features, we need to decide just how mismatching they are. For example, the political philosophies of Clinton and Reagan are very different. Their ages are pretty different (Reagan was the oldest president, and Clinton one of the youngest), though both are adults, and so the difference in age is not as large as it could be. However, their typical clothes are not so different. Clinton differs from Reagan a great deal on political beliefs, then, moderately on age, and only a little on clothing.

In order to calculate similarity, we need to quantify both of these factors: the importance of the dimension and the amount of similarity on a given dimension. Medin and Schaffer suggested that the amount of mismatch of each feature should be given a number between 0 and 1. If two items have matching features, they get a 1; if they have mismatching features, they get a number that is lower, with 0 indicating the greatest possible difference on that dimension (this score is typically not given, for reasons that will become apparent). However, the importance of the dimension would be represented by raising or lowering that number. For example, suppose that Reagan's favorite color is yellow, and Clinton's is blue. These are very different colors, but this dimension is not very important. Therefore, Medin and Schaffer (1978, p. 212) suggest that we could still give this item a difference score near to 1. By not paying attention to a difference, you are effectively acting as if the items have the same value on that dimension. So, Medin and Schaffer's rule combines the importance of the dimension and the difference on that dimension in one score.

How similar are Clinton and Reagan, then? If we were to use a similarity rule like the one used by the prototype model described above, we would take the mismatch score of each feature and add it up for all the features (hence, an additive rule). Things that are similar would have many matches and so would have higher scores. Medin and Schaffer, however, suggested that we should multiply together the scores

for each feature. Because the mismatch scores range between 0 and 1, a few mismatches will make overall similarity quite low. For example, suppose that Reagan and Clinton were identical on 25 dimensions but differed on 3 dimensions. And suppose that their differences on those dimensions were moderate, so that they received mismatch scores of .5 each. The overall similarity of Clinton and Reagan would now be only .125 on a 0–1 scale (25 scores of $1 \times .5 \times .5 \times .5$ for the three mismatching dimensions). This does not seem to be a very high number for two items that are identical on 25 out of 28 dimensions. If you work through a few examples, you will see that any mismatching feature has a large effect on diminishing similarity when a multiplicative rule is used. That is, mismatching features actively lower similarity on a multiplicative rule, but they simply do not improve it on an additive view. (You can now see why a similarity score of 0 is seldom given. If a dimension were given a 0, then the entire similarity of the two items is 0, since 0 multiplied by any number is always 0. That single mismatch would result in the two items being as different as possible.)

So far, we have been discussing how to decide the similarity of an item to a single other item. But what if you are comparing an item to a set of exemplars in a category? If you see an animal run across the road, and you want to decide whether to apply the brakes to avoid hitting it with your car, you might first try to identify what kind of animal it is (no point in risking an accident for a squirrel, you might feel) by comparing it to exemplars of other animals you have seen. So, you would compare it to cat exemplars, dog exemplars, raccoon exemplars, skunk exemplars, possum exemplars, and so on. Medin and Schaffer (1978) suggested that you add up the similarity scores for each exemplar in a category. So, if you have 150 exemplars of dogs in memory, you add up the similarities of the observed animal to these 150 items, and this constitutes the evidence for the animal being a dog. If you have 25 possums in memory, you add up the 25 similarity scores to get the evidence for possums, and so on. Loosely speaking, the category with the most similarity to the item will "win" this contest. If the item is similar to a number of categories, then you will choose each category with a probability that is proportional to the amount of similarity it has relative to the others. So, you might decide that such an animal is a raccoon on 60% of such cases but a skunk on 40% of the cases. For mathematical details, see Medin and Schaffer (1978) or Nosofsky (1984).

Because of the multiplicative nature of the similarity score, the Medin and Schaffer rule has an important property: It is best to be *very* similar to a few items, and it is unhelpful to be somewhat similar to many items. Imagine for example, that this animal was extremely similar to two other dogs you have seen, but not at all similar to any other dogs. The animal will have high similarity to these two items (near to 1.0), which would make the dog category have an overall similarity of almost 2.0. Now imagine that the animal shared three features with every possum you know (25 of them) and was different on three features with every possum. If the three mismatching features each have mismatch values of .25, then the animal would have a similarity of $1 \times 1 \times 1 \times .25 \times .25 \times .25$ to each possum, which is only .015625. When this is added up across all 25 possums, the total similarity for the category is only about .39—considerably less than the 2.0 for the dogs. Perhaps counter-intuitively, the two highly similar dogs beat out the 25 somewhat similar possums. This is because the mismatching features result in very low similarities, which cannot easily be made up by having many such examples. In short, *the Medin and Schaffer model says that it is better to have high overlap with a few items than to have moderate overlap with many items*. This property turns out to be a critical one, as the next chapter will reveal.[2]

Prototype advantages. Some empirical results initially discouraged researchers from seriously considering exemplar approaches. One such result was the prototype advantage that was found by Posner and Keele (1968, 1970) and others. When subjects learned categories by viewing distortions of a prototype (see above), they were often better at categorizing the prototype than another new item. (Though they were *not* initially better at identifying the prototype than the old items, as is commonly repeated.) This suggested to many researchers that subjects had actively abstracted the prototype from seeing the distortions. That is, they had learned what the items had in common and stored this as a representation of the whole category. (Because Posner and Keele 1968, actually started with a prototype to make the category members, it is natural to think of subjects who are exposed to the category members doing the same process in reverse.) But from the above discussion, you can see that an exemplar model could easily explain this result. Perhaps the prototype is more similar to the learned exemplars than a new, nonprototypical item is. Since the prototype was the basis for making all the other items, it must be similar to *all* of them to some degree, but this is not true for an arbitrary new item.

Another result was more of a problem for the exemplar view. Posner and Keele (1970) examined categorization for items immediately after category learning and after a one-week delay. When tested immediately, subjects were most accurate for the specific items that they had been trained on (old items), next most accurate for the prototype, and less accurate for new items they had not seen in training. But when tested a week later, memory for the old items declined precipitously, whereas the prototype suffered only a slight decrement. Posner and Keele argued that if the

prototype advantage were due to memory for old exemplars, it should have shown the same kind of decrement after a delay. They concluded that subjects formed a prototype during learning, and this prototype is somehow more insulated against memory loss than are memories for individual exemplars (perhaps because the prototype is based on many presented items, not just one). A similar effect was reported by Strange, Keeney, Kessel, and Jenkins (1970), and Bomba and Siqueland (1983) found the same effect in infants. There are other variables that have similar effects. For example, as more and more exemplars are learned for a category, the memory for old items decreases, but the prototype advantage generally increases (e.g., Knapp and Anderson 1984). If prototype performance is caused by memory for specific exemplars, how could performance on the prototype and old exemplars go in different directions?

An explanation for this kind of result was given by Medin and Schaffer (1978). First, consider why it is that in many cases with immediate testing, old exemplars are remembered best of all. This must be because when the item is presented at test, it results in the retrieval of itself. That is, when item 9 is presented at test, you remember having seen it before, and this makes you particularly fast and accurate at categorizing it. If you forgot this particular item, then it would no longer have this good performance, because it must be less similar to any other item than it is to itself. Over time, however, this loss of memory is just what happens. So, perhaps item 9 was a red circle over a green square. Even though you learned this item during the first part of an experiment, after a week's delay, this memory would be degraded. Perhaps you would only remember that it was a red circle over something green. Now when you get item 9 at test, it is not so very similar to its own memory, because that memory has changed. Similarly, if you learned 25 items as being in category A, your memory for each individual item will not be as good as if you only learned 4 items in this category. As you learn more and more, there is interference among the items which causes the exemplar memory for any one of them to be less accurate. This explains the decrements obtained in performance on old items.

What happens when you present the prototype (which was never seen during learning) at test? If it is an immediate test, the prototype is fairly similar to many items in the category, and so it is easy to categorize. However, it is not identical to any individual item, and so it is still not as fast as the old exemplars are. So, the advantage of old items over prototypes on immediate test is, according to the exemplar model, caused by the fact that the former has a perfect match in memory but the latter does not. (Note that this explanation relies on the exemplar model's weighting close similarity to one item more than moderate similarity to many items.) At delayed testing, the prototype is still very similar to a number of items. In fact,

Calculating Similarity According to the Context Model

Imagine that you have been learning categories made up of geometric figures printed on cards. Each figure is defined by shape, color, size, and position (left or right on the card). After you've learned the categories, suppose you're shown a new item, a large, green triangle on the left of the card. How do you calculate its similarity to the other items in order to decide its membership? The discussion in the main text gives the general outline of how this would be done. This box describes in a bit more detail how this is actually calculated in experiments that attempt to derive exact predictions for the context model.

Given the large green triangle on the left, one would have to compare it to each remembered exemplar. Suppose that you also remember seeing a large blue triangle on the right. We need to decide the matching and mismatching value for each dimension to see how similar it is. The two stimuli match on two dimensions, size and shape, and so these will be given values of 1.0. The two stimuli mismatch on two dimensions, and so these will be given values of s_c (for color) and s_p (for position). The s_c indicates how similar the green of one stimulus is to the blue of the other stimulus. If these are considered fairly similar, the value will be close to 1; if they are considered rather different, then the value would be closer to 0. The s_p correspondingly indicates how similar the left and right positions are. By using the multiplicative rule, we can calculate the entire similarity of these two stimuli as $1 \times 1 \times s_p \times s_c$. The problem is, how do we know exactly what s_p and s_c are so that we can come up with an actual number? In general, the answer is that these numbers will be calculated from the results of the experiment itself. For example, we can see how likely people are to categorize an item that is just like a learned item but differs in color; and we can see how likely people are to categorize an item that is just like a learned item but differs in shape; and so on. By using a mathematical modeling program, researchers in this area can put in the expected formulas for each item (i.e., how similar each test item is to each learned item according to the multiplication rule), and the program will provide the values of s_p, s_c, and the other similarities that make the model perform as well as possible. These are called *free parameters* of a model, because they are estimated from the data, rather than being stated by the theory in advance. (Other theories also have free parameters. For example, I said that prototype theory often has weights on how important each feature is for a category. These could be estimated from the data as free parameters, though they could also be directly measured through means analogous to those described in the next paragraph.)

Unfortunately, this is not entirely the end. Recall that Medin and Schaffer also discussed the possibility that some dimensions might be attended to more than others. Suppose, for example, that subjects never paid attention to the position of the figures for some reason, perhaps not thinking that it was relevant. Now, the value for s_p that we calculate by the above procedure would include *both* the intrinsic similarity of the items and the attention that subjects give to it. If subjects really ignored position, then s_p would equal 1—suggesting that the right and left positions were viewed as identical. That is, there is no way to separate the mismatch score from the amount of attention

continued

that people pay to that dimension, because both are being used by subjects to make categorization decisions. This is not necessarily a problem, but sometimes one does wish to know what the real similarities are, separately from knowing which dimensions subjects paid more attention to.

One way to handle this concern is to choose the stimulus differences so that they are known to be equally different in different dimensions. For example, by asking different subjects to provide similarity ratings, it might be possible to choose a size difference (values for the large and small items) that is psychologically the same size as the color difference (between the particular values of blue and green), which is also equal to the shape difference, and so on. The experimenter can ask subjects to rate the similarity of all the stimuli before doing a categorization experiment. Then he or she can choose stimulus values that are equated, or at least the relative values of s_p, s_c, and the rest can be measured. Unfortunately, this technique cannot tell us how much attention people pay to each dimension during learning—only how similar the stimulus values are perceptually. To discover any attentional differences, one must still estimate the s parameters from the main experiment.

Calculating the exact predictions for these models is not something that can easily be done with paper and pencil, as can be seen by this description. The actual calculations of the s values and therefore the model's precise predictions is almost always done in conjunction with a mathematical modeling program. In other cases, the properties of the models can be illuminated by proofs using the general form of the similarity rule (e.g., Nosofsky 1984; 1992). But these are not tasks for beginners.

because some of the specific information about the old items has been lost, they may now seem even *more* similar to the prototype than they did before. For example, suppose that the prototype was a red circle over a green triangle. This matched item 9 to a large degree, since the two differed only in the shape of the bottom figure. After forgetting, though, the memory for item 9 actually matches the prototype more than before, because there is no longer a conflict between the square in item 9, which has been forgotten, and the triangle in the prototype. In short, the effect of forgetting is to make the old test items less similar to their own memories, but it has less effect (or even the opposite effect) on the prototype.

Hintzman and Ludlam (1980) performed simulations of an exemplar model showing that the above explanation could in fact explain why exemplar memory and prototype memory have separate time courses. They showed that as forgetting occurred, old items became steadily less similar to the remembered exemplars, but prototypes retained their similarity to exemplars for the most part. (See also Hintzman 1986, for a more complete modeling effort.) In short, even if performance is based only on remembered exemplars, prototype effects will occur because prototypes are quite similar to a number of exemplars. One does not need to propose a summary representation of a category in order to explain prototype effects.

What Is an Exemplar?

For some time now, I have been discussing "exemplars" and the view of concepts based on them. However, I have not yet precisely said what these entities are. To some degree, this question has been left open by the proponents of this view. For example, suppose that I see a squirrel run across a lawn while I walk in to work today. Does this brief glimpse of a squirrel constitute an exemplar, even if I don't pay much attention to it? I have seen hundreds, perhaps thousands of squirrels in this way. Are all these exemplars stored in memory? (And are they stored *as* squirrels? As explained above, encoding the exemplar's category is required for it to influence categorization.) The exemplar view of concepts does not necessarily have an answer to this question, which is in part an issue of memory in general (not just concepts). It would have to say that *if* I did notice and remember that squirrel, then it could have an effect on the way I identify later animals as squirrels. If the squirrel is not encoded into memory or is forgotten, then it clearly can have no effect. But a theory of concepts cannot say exactly what everyone will and will not remember.

There is another, deeper question about exemplars, namely how an exemplar is to be defined. Consider this example. Suppose that I know a bulldog that drools a great deal named Wilbur. In fact, this bulldog lives next door to me, and so I have many opportunities to see him drool. I have seen other bulldogs, some of which appear to be drooling, and some of which do not. How do I decide, now, whether a friend of mine, who is complaining about her new dog's drooling, has a bulldog? According to the exemplar view, I would have to retrieve all the dog exemplars that I know that drool (no small number), and then essentially count up how many of them are bulldogs. But in retrieving these exemplars, how do I count Wilbur? Does he count once, because he is only one dog, or does each *encounter* with Wilbur count separately? Put in more formal terms, do I count types (Wilbur) or tokens (Wilbur-encounters)?

In terms of making an accurate decision, it seems clear that I should only count Wilbur as a type—he should only count as one dog. If I count up every Wilbur encounter as another exemplar, then the fact that a (drooling) bulldog lives next to me is having a large effect on my decision; if a labrador lived next to me, I would have many, many fewer such exemplars. However, which kind of dog happens to be living next door is not relevant to the question of whether a drooling dog is a bulldog. Or, to put it another way, how often I happen to see one particular bulldog should not greatly influence my decisions about bulldogs in general.

Nosofsky (1988) addressed this question in an experiment using colored patches as stimuli. The patches varied in how saturated and bright the colors were: The more saturated and bright colors tended to be in one category, and the less saturated and bright colors in another. He varied the frequency with which items were presented: One of the items was presented five times as often as the other items during learning. If each exemplar is considered as a type, then this frequency manipulation should not influence later category decisions. The fact that one color keeps reappearing would be like the fact that I live next door to Wilbur, not a relevant indication of the category in general. But if exemplars are defined as tokens, then stimuli that were like the more frequent item would be better category examples than stimuli that were like other, less frequent items, because there would be "more exemplars" remembered for the more frequent one. This is exactly what Nosofsky found. After learning, he showed subjects the items and asked them to rate their typicality. The more frequent item and other items that were close to it were rated as being more typical than the less frequent items. By this result, then, an exemplar is not an actual thing but rather the encounter with a thing. So, if I encounter Wilbur a hundred times, this creates 100 exemplars, not just one. Nosofsky also provided a simulation of the exemplar model, showing that it accounted for the results much better if it considered each presentation of the stimulus as an exemplar, rather than each type as being an exemplar.

Barsalou, Huttenlocher, and Lamberts (1998) raised a possible problem with this interpretation of Nosofsky's experiment. They pointed out that we do not know what subjects thought about the reappearing colors. Perhaps they thought that the stimuli were somehow different objects even if they looked identical. (It is difficult to know how to interpret the reappearance of these items, since they were color patches rather than objects.) Furthermore, as color patches are difficult to remember precisely, perhaps people did not realize that exactly the same item was being shown so often. Perhaps they thought that the colors were slightly different. If so, then they would naturally count them as separate exemplars.

Barsalou et al. performed a clever experiment in which they showed two groups of subjects the exact same stimuli during learning, but they varied whether subjects thought that each stimulus was unique, or whether they were seeing some of the items multiple times. Under most conditions, they found that this manipulation had virtually no effect on the concepts people formed; the very frequent exemplar had a strong effect in both conditions. That is, to return to my example, it makes no difference whether I think I'm seeing 100 different bulldogs or one bulldog 100 times—the effect on my concept of bulldogs is the same.[3] As Barsalou et al. point out, this has implications for prototype models as well as for exemplar models of concepts. In both cases, the theory needs to specify how units are counted up (how many features/exemplars have been viewed), and the empirical results suggest that it is *encounters* with objects that are most important, rather than the objects themselves.

The Knowledge Approach

The discussion of the final major theory is actually a bit premature for the present chapter. The prototype and exemplar models arose from the ashes of the classical view, and they were designed to account for the data that were so problematic for that view. The *knowledge approach* in contrast arose as a reaction to the two other approaches, and it is in some sense built upon them. As a result, we have not yet discussed the experimental results that led to this view, nor are we ready to do so now. A later chapter (chapter 6) provides a more detailed exposition of this account. Nonetheless, it will be useful to have this approach in mind when reading the next few chapters, and so I will give a somewhat brief description of this view now, without providing much of the experimental evidence.

The knowledge approach argues that concepts are part of our general knowledge about the world. We do not learn concepts in isolation from everything else (as is the case in many psychology experiments); rather, we learn them as part of our overall understanding of the world around us. When we learn concepts about animals, this information is integrated with our general knowledge about biology, about behavior, and other relevant domains (perhaps cuisine, ecology, climate, and so on). This relation works both ways: Concepts are influenced by what we already know, but a new concept can also effect a change in our general knowledge. Thus, if you learn a surprising fact about a new kind of animal, this could change what you thought about biology in general (e.g., if you learn that snails are hermaphrodites, your knowledge about sexual reproduction in general could be affected); and if something you learn about a new animal doesn't fit with your general knowledge, you may have cause to question it or to give it less weight. (Could snails *really* be hermaphrodites? Maybe you just misunderstood. Best to say nothing and hope they go away.) In general, then, the knowledge approach emphasizes that concepts are part and parcel of your general knowledge of the world, and so there is pressure for concepts to be consistent with whatever else you know (Keil 1989; Murphy and Medin 1985). In order to maintain such consistency, part of categorization and

other conceptual processes may be a reasoning process that infers properties or constructs explanations from general knowledge.

Let me just give a simple example of the kind of knowledge involved here. One area of knowledge that is often studied in research from this perspective is that of biology. We know things about evolution, reproduction, physiology, and ecology, part of which we have just learned "naively" (on our own or informally from parents), and part of which we learned through more formal study. However, even young children seem to have basic ideas about biology (Gelman and Wellman 1991; Keil 1989) that they use in making judgments of the following sort. If a child sees a fuzzy, gray, tiny animal paddling around after a large, white, goose, the child may conclude that the animal must be a goose as well, even though it looks very different from other geese it has seen. Apparently, the child is using the logic: "Babies are smaller than their parents, and they often stick close to their parents. Any baby of a goose must itself be a goose. So, this much smaller animal could well be a baby, even though it looks rather different from the goose, and so it is also a goose." Of course, the child doesn't say this out loud, but there is reason to think that children are sensitive to notions of inheritance and parentage—basic biological properties that then influence their categorizations. In general, this approach says that people use their prior knowledge to reason about an example in order to decide what category it is, or in order to learn a new category.

In one description, this aspect of concepts was referred to as "mental theories about the world" (Murphy and Medin 1985), which is accurate enough if one understands that people's naive theories are incomplete and in some cases contradictory, given our incomplete knowledge and understanding of the things around us. The child in the above example doesn't have a complete theory of biology but does know some basic facts and principles that are partly integrated. Thus, this approach is sometimes called the *theory view* (or even the *theory theory* by those less easily embarrassed than I am). However, the term *theory* suggests to many something more like an official scientific theory, which is probably not an accurate description of people's knowledge (see, e.g., Gentner and Stevens 1983). This has caused some confusion about exactly what the approach is claiming, so I will tend to use the term *knowledge* rather than *theory*, to avoid this potential confusion.

Some of the discussion of schemata discussed in the prototype view is relevant here as well. For example, one reason given for using schemata for representing concepts was that they can represent relations between features and dimensions. This is just one way of representing knowledge about the domain. For example, we may know that animals without wings cannot fly, and so there may be a relation between the schema slot describing body parts and the slot describing behaviors that manifests this relation.

One of the studies on typicality described in some detail in the previous chapter was also a motivation for the knowledge view, namely, Barsalou (1985). Recall that Barsalou found that *ideals* are important to determining typicality. For example, something might be considered a good example of a weapon to the degree that it was an efficient way to hurt or kill people. This "ideal" weapon is not the average value of all weapons, because most weapons are less than ideal on this account (e.g., a knife requires close distance, accurate handling, and can only cut one person at a time). Barsalou found that items that were closer to the ideal were more typical than items that were farther away, and this was true even when family resemblance was first factored into the typicality judgment. This influence of ideals cannot, then, reflect just pure observation of the category, as a prototype or exemplar approach might claim. If people relied on the weapons they had seen or heard about, they would find only moderately effective devices to be the most typical weapons. Similarly, they would expect only moderately efficient people-movers to be good vehicles, since on average, vehicles are by no means ideal people-movers.

Where do these ideals for categories come from, then? Most likely they come from our knowledge of how each category fits in with other parts of our lives—its place in our greater understanding of the world. We know that vehicles are made *so that* people can be moved from place to place. Therefore, the most typical vehicles would do this in the best possible way. We know that weapons are created in order to hurt (or threaten to hurt) other people. Therefore, the most typical ones are the ones that do this in an effective way. Furthermore, we can apply our general knowledge to evaluate how well different vehicles and weapons actually fulfill these functions.

The importance of such knowledge can be illustrated even more by a kind of category that Barsalou (1985) called *goal-derived categories*. These are categories that are defined solely in terms of how their members fulfill some desired goal or plan, for example, things to eat on a diet, things to take from one's house during a fire, good birthday presents, and so on. For goal-derived categories, very little of the category structure is explained by family resemblance. For example, things to eat on a diet might include celery, sugar-free jello, diet soda, baked potatoes, baked fish, and skim milk. These items differ in many respects. They are much less similar to one another than normal food categories such as dairy products or meats, yet, they are all within the same category by being things that people eat while on a diet. Here, the ideal is something like having the smallest number of calories or the least

fat. So, celery is an excellent example of things to eat on a diet, because it has virtually no fat and is extremely low in calories. Bread is a fairly good example, though it has somewhat more calories and fat. Fruit juice might be a moderate example, since it is low in fat but not particularly low in calories. And ice cream would be a bad example. Barsalou found that the most typical examples of goal-derived categories were the ones that were closest to the ideal. Family resemblance did not explain a significant portion of the variance. This is an extreme case in which an item's place in a larger knowledge structure is perhaps the most important aspect of category membership, and the "average" properties of the category members count for little. For example, the best food to eat on a diet would be a filling food with no calories, fat or other bad ingredients. However, these properties are by no means the most frequent ones of foods that people actually eat while on diets. So, this ideal seems to be imposed by our understanding of what the category is supposed to be, which is in turn driven by our understanding of how it fits into the rest of what we know about foods and their effects on our bodies. The ideal cannot be derived from just observing examples and noting the features that occur most. Although the goal-derived categories are an extreme example of this (because the members have very little in common besides the ideal), Barsalou found evidence for the importance of ideals in common categories as well (see previous chapter). Also, recall that Lynch et al. (2000) found a similar pattern for the category of trees; many other related examples will be described in chapter 6.

One of the themes of the knowledge approach, then, is that people do not rely on simple observation or feature learning in order to learn new concepts. They pay attention to the features that their prior knowledge says are the important ones. They may make inferences and add information that is not actually observed in the item itself. Their knowledge is used in an active way to shape what is learned and how that information is used after learning. This aspect of the theory will be expounded in greater detail in chapter 6.

One clear limitation of the knowledge approach should already be apparent: Much of a concept cannot be based on previous knowledge. For example, you might have previous knowledge that helps you to understand why an airplane has wings, how this relates to its ability to fly, what the jets or propellers do, and so on. However, it is probably only by actual observation of planes that you would learn where propellers are normally located, what shape the windows are, that the seats usually provide no lower back support, and so on, because these things are not predictable from your knowledge before learning the category. Or, in other cases, it is only after observing some category members that you know what knowledge is relevant and can only then use it to understand the category (see Murphy 2000, for a discussion). So, the knowledge approach does not attempt to explain all of concept acquisition by reference to general knowledge; it must also assume a learning mechanism that is based on experience. However, this approach has not incorporated any empirical learning mechanism. This may be seen as a shortcoming of the knowledge approach, or one can view the empirical learning process as simply being a different problem. That is, proponents of the knowledge approach are pointing out the ways that prior knowledge influences the learning of a new concept, and other aspects of learning are not part of the phenomena they are trying to explain. However, it should be clear that a complete account requires an integrated explanation of all aspects of concept learning. Furthermore, we should not necessarily assume that the empirical and knowledge-based learning components will be easily separable modules. It is possible that the two interact in a complex way so that one must study them together to understand either one (Wisniewski and Medin 1994). However, that discussion must await a later chapter.

Conclusions

It is too early in this book to evaluate these different approaches. However, it is worth emphasizing that none of them suffers from the problems of the classical approach. All of them actively predict that categories will have gradations of typicality and that there will be borderline cases. Unlike the later revisions of the classical model (discussed in the previous chapter; e.g., Armstrong, Gleitman, and Gleitman 1983), these theories claim category fuzziness as an integral part of conceptual processing, rather than an unhappy influence of something that is not the "true" concept. This is because similarity of items is inherently continuous. Category members will be more or less similar to one another and to their prototype, and this gradation of similarity leads to typicality differences, RT differences in judgments, and learning differences. Similarly, whether an item is consistent with one's knowledge in a complex domain is not an all-or-none matter but often a question of relative consistency. Thus, unlike for the classical view, typicality phenomena are not a nuisance to be explained away, but are rather inherent to the working of these approaches.

Another point to be made about each of these approaches is that they are not as entirely self-sufficient as one might like. For example, the prototype view does not deny that people learn and remember exemplars. Clearly, if Wilbur, the bulldog, lives next door to me, and I see him many times per week, I will be able to identify

him and his peculiar attributes. And, as already mentioned, the first time one encounters a category member, the only prototype one can form would be based on that single exemplar. Thus, exemplar knowledge and prototype knowledge must exist side by side to at least some degree, according to prototype theory. The general claim of that theory, however, is that for mature categories, people rely on summary representations of the entire category rather than specific exemplars in making judgments about the concept.

Similarly, I pointed out that the knowledge approach focuses on one important (it claims) aspect of learning and representing concepts. However, it must admit that there is an empirical learning component to concepts, if only to explain the results of psychological experiments that use artificial stimuli that are removed from any knowledge. It is likely, then, that this view will have to be combined with one of the other views in order to form a complete theory of concepts. Finally, exemplar theorists might also agree that there must be a level of general knowledge that is separate from exemplar knowledge and that affects concepts and their use (though in fact most have not addressed this issue). For example, one could argue that facts such as whales being mammals are school-learned general facts, rather than something one learns from seeing many whale exemplars. But of course one does see whale exemplars occasionally too. So, in answering questions about whales, some information might come from the exemplars and some from general knowledge.

This mixture of different kinds of conceptual knowledge makes it difficult to evaluate the different theories. The result in the field has been to focus on certain experimental paradigms in which such mixtures would be expected to be less likely. However, for real-life concepts, we would do best not to assume that a single form of conceptual representation will account for everything.

APPENDIX: THE GENERALIZED CONTEXT MODEL

The body of this chapter has discussed the Context Model, proposed by Medin and Schaffer (1978), as the most influential version of the exemplar approach. However, in more recent times, Robert Nosofsky's enhancement of this, the *Generalized Context Model* or *GCM*, has been more widely cited and tested. This model has a number of different forms, as it has developed over years of investigation. The purpose of this appendix is to describe it in more detail and to give an idea of how it works. For more formal descriptions, see Nosofsky (1992) or Nosofsky and Palmeri (1997).

Overview

Recall that exemplar models argue that you categorize objects by comparing them to remembered exemplars whose categories you have already encoded. The more similar the object is to exemplars in a given category, the more likely it is to be categorized into that category. The categorization process can be thought of as having three parts. First, one must calculate the distance between the exemplar of interest and all the other exemplars. In the Medin and Schaffer (1978) model, this was done by the multiplicative rule described in the main chapter. In the GCM, this is done by a more general distance metric. Second, this distance metric is scaled in a way that has the effect of weighting close similarity much more than moderate similarity. Third, once one has all these similarities of the object to known exemplars, one must decide which category the object is in. This involves a fairly simple comparison of the exemplar similarities of the different categories involved.

The formulas involved in the GCM can be difficult if you aren't familiar with them. I personally am not a math modeling expert, and it has taken me some time to understand them (to the degree that I do). This chapter will help you begin to understand the model. However, to reach a fuller appreciation, I would recommend that when you read individual papers that describe this or other models, you resist the very natural temptation to skip over all the formulas. If you read these formulas and the authors' explanations of them in four or five different papers, you will eventually find (perhaps to your amazement) that you understand what they are talking about. So, don't give up just because it all seems so complicated here. I describe the three parts of the GCM in the following sections.

Distance Calculations

The basic idea of exemplar models is that an object brings to mind other similar exemplars. To measure this, the model calculates the psychological distance between the object to be categorized (usually called *i* below) and all the known exemplars, in order to identify which exemplars are going to be brought to mind.

In order to compare various stimuli, the GCM assumes that they have been broken into dimensions of some kind. This could be because they are items that have been constructed according to certain dimensions (color, size, position on a card, etc.) that subjects would be sensitive to. In other cases, one may have psychological dimensions of the stimuli that are derived from a scaling program, such as multidimensional scaling. For example, using multidimensional scaling, Rips, Shoben, and Smith (1973) discovered that animals were thought to be similar based on their size and predacity. The scaling program provided the values of each item on each dimension, and these would be used to calculate similarity. The distance between any two objects is a function of how far apart they are on each dimension.

Equation (1) shows how the GCM calculates the distance between two items, i and j, when the dimensions of the objects are known. (Think of i as being the object to be categorized, and j one of the remembered exemplars.) It is essentially the same as the Euclidean distance between two points in space, from high-school geometry.

$$d_{ij} = \sqrt{\sum_{m} w_m |x_{im} - x_{jm}|^2} \tag{1}$$

For each dimension m of the items (either a physical dimension or one from a multidimensional scaling solution), you calculate the difference between the items $(x_i - x_j)$ and square it. Then this number is multiplied by the weight for that dimension (w_m). So, important dimensions will be counted more heavily than less important dimensions. (In the original context model, this was reflected in the mismatch values. The GCM more clearly separates psychological similarity from dimensional importance in categorization.) These numbers are added up for all the m dimensions, and the square root is taken. The main difference between this and simple Euclidean distance is the weighting of the dimensions. In calculating real distances, no spatial dimension is weighted more than any other. In the GCM, the w values for each dimension are a free parameter—that is, they are calculated from the data themselves rather than being specified by the model. Kruschke's (1992) influential ALCOVE model (based on the GCM to a large degree) provides a connectionist mechanism that actually learns which dimensions are important.

I should note that this particular form of the distance metric can vary in different studies. (Those who are just trying to get the basic idea of this model should definitely skip this paragraph.) Note that we squared the differences on each dimension and then took the square root of the sum. That is, we raised the distance to the power 2 and then raised the sum to the power 1/2 (that's a square root). Distance metrics in general raise the separation by some power and then the sum to one over that power. The number 2 is only one such number that could be used. We could have used the number 1—that is, raise the distances to the power of 1 (i.e., do nothing) and taken the sum to the power of 1/1 (i.e., do nothing again). This would give us the "city block" metric, in which the distance between two points is the sum of their distances on each dimension. This is called the city block metric, because one does not form hypotenuses for distances—one cannot cut through a city block.

To get from 40th Street and Second Avenue to 43rd Street and Sixth Avenue, one must walk three blocks north and four blocks west, resulting in seven total blocks (let's assume the blocks are squares). By Euclidean distance, the shortest distance connecting these points (cutting across) would be only five blocks. For some stimulus dimensions, the city block metric seems most appropriate, whereas for others, Euclidean distance works best.[4] And for still others, some number in between is most appropriate. Thus, this difference between dimensions can be made into a variable (usually called Minkowski's *r*), which is altered depending on the nature of the stimuli. So, you may see something like equation (1) with *r*s and 1/*r*s in it.

Turning Distances into Similarity

Unfortunately, we are still not done with deciding the psychological distance between two items. Research has shown that behavioral similarity between items is an exponentially decreasing function of their psychological distance (Shepard 1987). For example, if rats learn to produce a response to a certain colored light, other lights will elicit the same response as a function of the exponential distance between the test light and the original. The exponential function has the effect that things that are extremely close to the object have a large effect, which falls off very quickly as things become moderately and less similar. Recall that Medin and Schaffer (1978) used a multiplicative rule so that objects that are moderately different would have little effect on categorization decisions. The exponential function does the same thing.

As shown in equation (2), the distance scores derived from equation (1) are input to the exponential function, producing a similarity score, s.

$$s_{ij} = \exp(-c \cdot d_{ij}) \tag{2}$$

Note that $\exp(x)$ means to raise e to the xth power, e^x, and that $\exp(-x) = 1/e^x$. Therefore, the bigger x gets, the smaller $\exp(-x)$ gets. In equation (2), this means that the greater the distance between i and j, the smaller the similarity, since the distance is being negated. When distance is 0 (i.e., two stimuli are identical), s would equal 1.0; otherwise, s falls between 0 and 1. The variable c basically determines the spread of similarity by modulating the effect of distance. Sometimes people seem to pay attention to only very close similarity, and other times people take into account even fairly weak similarity. A high value of c corresponds to the former situation, and a low value corresponds to the latter. If c is very high, then the exemplar model is essentially requiring that an item be identical to a known exemplar, because any distance between i and j would be multiplied by a high number, resulting in a low similarity. If c is very low, then the similarity to all known items is used. Usually, c is a free parameter estimated from the data.

Making the Decision

In order to make a categorization decision, one has to decide which category exemplars are most like the object being categorized, i. If the object is similar to many dogs, a few cats, and one burro, then it is probably a dog. The GCM does this by looking at the object's similarity to *all* objects in every category, and then comparing the similarity of one category to that of all the others. In the previous sections, we calculated the similarity of the object to every other exemplar. If you were to add up all these similarity scores, you would know the total pool of similarity this object has to everything. The GCM uses something called the Luce Choice Axiom to turn this similarity into a response. The Luce Choice Axiom basically asks how much of the total pool of similarity comes from dogs, how much from cats, and how much from burros, and then turns those answers into response probabilities. Equation (3) calculates the probability that the object, i, will be placed into each category, J.

$$P(J \mid i) = \sum_{j \in J} s_{ij} \Big/ \left[\sum_{K} \sum_{k \in K} s_{ik}\right] \tag{3}$$

The numerator is the similarity of i to all the members j of category J (as calculated in equation (2)). The more similar the item is to known exemplars in J, the higher this probability. The denominator is the similarity of i to members of *all* known categories (K), the total pool of similarity mentioned above. In English, then, equation (3) is the ratio of the similarity of i to all the things in J to the total similarity pool.[5] Continuing the previous example, if i is mostly similar to dogs, then perhaps $P(\text{dog} \mid i) = .75$, because dogs have 75% of the total similarity for this object. And perhaps $P(\text{cat} \mid i) = .20$, because the object is similar to a few cats. And because you once saw a very strange burro, $P(\text{burro} \mid i) = .05$. What the Luce Choice Axiom says is that you should call this object a dog 75% of the time, a cat 20% of the time, and a burro 5% of the time.

Note that this formula is probabilistic. It doesn't say that people will always categorize i into the category with the highest score. Instead, it says that their categorization will match the probability score. This behavior is not entirely rational. If it is 75 percent likely that i is a dog, then I should obviously categorize it as a dog whenever I see it, because it is much more likely to be a dog than a cat or a burro, based on my own experience. Any other response is much less likely. However, people tend to do *probability matching* in these situations, in which they give less likely responses a proportional number of times, rather than always choosing the most likely answer. This behavior is found in many choice tasks in both humans and other animals.

Later versions of the GCM have sometimes incorporated another parameter proposed by Ashby and Maddox (1993), called *gamma*, which relates to the probability matching phenomenon. To do this, the numerator of (3) is raised to the gamma power, and the inside term of the denominator ($\sum_{k \in K} s_{ik}$) is also raised to that power. (The exact value of gamma is usually a free parameter.) That is, once one has calculated the total similarity of i to a category, that similarity is raised to the power gamma. Gamma has the effect of varying the amount of determinacy in subjects' responses. When gamma is 1, the categorization rule has the characteristics mentioned earlier: Subjects respond proportionally to the probability. When gamma is higher, they respond more deterministically: They tend to choose the most likely category more. This is because the similarity values are less than 1, so raising them to a high power tends to decrease the small values almost to 0, thereby benefiting the larger values.

See how simple it all is? No? Well, you are not alone. Although all this may be confusing on first reading, if one simply focuses on one variable or equation at a time, one can usually understand how it works. Attempting to keep the entire GCM in one's head at a time is pretty much impossible, so do not be too discouraged if you don't feel that you get it. Especially at the beginning, you can probably only understand it piece by piece.

Although the GCM has been extremely successful in modeling results of categorization experiments, one criticism of it has been that it is too powerful. The complexity of the model and the number of parameters it has makes it very good at fitting data, even if the data are exactly what would be predicted by a prototype model (Smith and Minda 2000). Perhaps because it is a rather late addition to the model, gamma has struck some in the field as being a dubious construct. Some critics feel that it is simply a free parameter that lets the GCM account for data that could not previously be accounted for, without any clear psychological evidence that determinacy of responding is a variable that changes systematically. However, this level of argument is far beyond that of the present discussion. See the interchange between Smith and Minda (2000) and Nosofsky (2000) for discussion and references.

Whatever the criticisms of it, the GCM has been an extremely influential categorization model. What should now be obvious is that there is a lot more to the GCM than simply saying that people remember exemplars. There are many assumptions about how similarity is calculated, how decisions are made, and what variables affect performance that go far beyond the simple claim of exemplar representations of concepts. Thus, the model's successes and failures cannot be taken as direct evidence for and against exemplars alone.