1

Similarity, typicality, and categorization

LANCE J. RIPS

Here is a simple and appealing idea about the way people decide whether an object belongs to a category: The object is a member of the category if it is sufficiently similar to known category members. To put this in more cognitive terms, if you want to know whether an object is a category member, start with a representation of the object and a representation of the potential category. Then determine the similarity of the object representation to the category representation. If this similarity value is high enough, then the object belongs to the category; otherwise, it does not. For example, suppose you come across a white three-dimensional object with an elliptical profile; or suppose you read or hear a description like the one I just gave you. You can calculate a measure of the similarity between your mental representation of this object and your prior representation of categories it might fit into. Depending on the outcome of this calculation, you might decide that similarity warrants calling the object an egg, perhaps, or a turnip or a Christmas ornament.

This simple picture of categorizing seems intuitively right, especially in the context of pattern recognition. A specific egg – one you have never seen before – looks a lot like other eggs. It certainly looks more like eggs than it looks like members of most other categories. And so it is hard to escape the conclusion that something about this resemblance *makes* it an egg or, at least, makes us think it's one. In much the same way, if you happen to be a subject in a concept-learning experiment and are told that your job is to decide on each trial whether a meaningless pattern of dots is a member of Category A or of Category B, then you might be right to think that resemblance must be the key to the correct answer. You may have nothing else to go on. Or again, in the case of a child learning what things should be called *egg*, it seems very likely that perceptual similarity to previously labeled eggs – and perhaps perceptual *dis*similarity to certain kinds of noneggs – will play a big role in the developmental story. (Even for artificial

categories and even for children's classification, similarity might not be the *whole* story; we shall return to this later in discussing results due to Carey, 1982; Fried & Holyoak, 1984; and Keil, 1986.)

But despite the intuitive appeal of this *resemblance* approach to categorizing, it has lately come in for some criticism. This criticism stems from philosophical discussions of similarity, particularly by Goodman (1970) and Quine (1969), but these antiresemblance views are now gaining ground in psychology too (Murphy & Medin, 1985; Oden & Lopes, 1982). For example, in a recent *Psychological Review* paper, Murphy and Medin (1985) argue that similarity is just too loose a notion to explain categorization adequately. Following Goodman (1970), they assert that similarity is highly relative and context-dependent: Our judgment of what is similar to what, according to this view, depends on the kinds of objects, properties, relations, and categories that we happen to have learned. But if that is right, then psychological similarity may depend on categorization rather than the other way around. In other words, the resemblance theory of categorization is vacuous unless we can specify how similarity is determined without begging the very questions about categorization that it is supposed to solve. But, in Murphy and Medin's view, the prospect for this is dim. They advocate an alternative approach in which concepts are taken to be minitheories about the nature of the categories they describe. Categorizing an object is then a matter of applying the relevant theory. So, on this view, everyday classification is very much like scientific classification, with the proviso that lay theories are likely to be less accurate and less detailed than those of scientists.

If this conclusion is correct, it has important implications, since it undercuts a whole class of models in cognitive psychology. In general, I think their conclusion *is* right, but I would like to try in this discussion to argue for it from a different angle. One reason for doing this is that you might question whether similarity is really as relative as Murphy and Medin make it out. True, there is evidence that similarity ratings in psychological experiments depend on context; they depend on the set of things that the subject is rating (Tversky, 1977; Tversky & Gati, 1978). But is this variation in ratings due to a change in our sense of similarity, or is it due instead to the constraints of translating this sense into a response on a one-dimensional rating scale? Murphy and Medin's arguments (as well as the arguments of their philosophical sources) have a 1950s-style New Look about them – things resemble each other because you believe they are in the same category – a view few psychologists these days wholeheartedly endorse. Second, you might also feel that the choice of theory-based models over similarity-based models is not as clear-cut as Murphy and Medin believe.

After all, does the notion of a theory really provide a firmer foundation for everyday categories? Even if similarity is as vague and variable as they claim, surely theories – especially lay theories – are not noticeably less vague or less variable. On first glance, the two types of models don't present much to choose from. Why not stick with similarity, then, where at least we have some inkling of the shape that a model would take, thanks to work by Tversky (1977) and others?[1]

In any event, what I would like to argue is this: Even if we grant that people have a stable sense of resemblance or similarity and even if we can give this sense a correct psychological description, similarity still will not be either necessary or sufficient for dealing with all object categories. As long as we stick to the ordinary meaning of similarity – the meaning that it has for nonexperts – then similarity will not be enough to explain human concepts and categories. To convince you of this, I'll present evidence from a set of experiments in which subjects were asked either to categorize an instance or to judge its similarity with respect to two potential categories. These experiments demonstrate that the favored response sometimes differs in the two tasks; that is, subjects may judge the instance more similar to Category A than to Category B but also judge the same instance more likely to be a member of Category B than Category A. On the basis of these results, I'll claim that there are factors that affect categorization but not similarity and other factors that affect similarity but not categorization. In other words, if all this goes through, there is a "double dissociation" between categorization and similarity, proving that one cannot be reduced to the other.

I'll begin by reminding you of the role that similarity plays in some well-known psychological theories of categorization, since that should make it easier to see exactly what damage would be done by undermining the resemblance theory. With this as background, I'll then describe two experiments that purport to show that similarity cannot be all there is to categorizing. That's the easy part. I'll then mention two more experiments that try to show that, for some concepts, similarity and categorization are actually independent. That's the hard part. A word of warning: None of these experiments is very high-tech. They all rely on simple ratings collected from groups of subjects. Some of them also make use of rather bizarre categories as stimuli. I'll try to head off some objections on this score in finishing up.

The role of similarity in theories of categorization

To start out, let's take a look at how similarity enters into models of categorization. Figure 1.1 lists several kinds of models, with those at

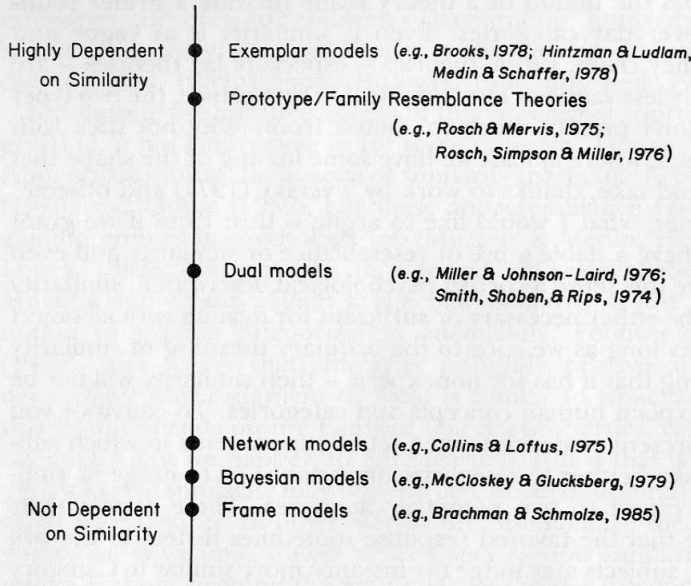

Figure 1.1. A summary of major approaches to categorization, arranged according to their theoretical dependence on resemblance.

the top being most dependent on similarity and those at the bottom least dependent. This list of models is far from complete, of course, and you should take it only as a very rough guide to the categorization terrain. The part that similarity plays clearly depends on what the theory takes to be the mental representation of categories; and a good rule of thumb might be that the more concrete the representation, the more dependent the theory is on similarity. In general, models at the top of the continuum assume fairly concrete representations, whereas those at the bottom are much more abstract. Beginning at the more dependent end, we have those theories that represent categories as specific instances. According to this type of approach, you mentally represent the category of eggs, say, in terms of memories of specific eggs that you have actually encountered. These exemplar theories, as they are usually called, are largely at the mercy of similarity. That is because, if your only source of knowledge about a category is a record of the exemplars you have seen, then the only way for you to decide if a new instance is also a member is to compare that instance to the remembered ones.[2]

A good example of the exemplar approach is the model proposed

by Medin and Schaffer (1978), which they call the "context theory." According to Medin and Schaffer, if you have to decide whether a given instance is a member of one of several categories, you do it by retrieving from memory a category member that the target instance reminds you of. The member that you retrieve then determines the category of the new instance. For example, if the target instance happens to remind you of a particular egg – maybe the one you had for breakfast today – then you will classify this new instance as an egg too. For our purposes, the important point is that the instance you are reminded of is assumed to be entirely a function of similarity. Medin and Schaffer put it like this: "The probability of classifying exemplar i into category j is an increasing function of the similarity of exemplar i to stored category j exemplars and a decreasing function of the similarity of exemplar i to stored exemplars associated with alternative categories" (Medin & Schaffer, 1978, p. 211). This reliance on similarity is also apparent in other instance-based approaches – for example, in the MINERVA model of Hintzman and Ludlam (1980) and in work of members of the Canadian School of Nonanalytic Psychology (e.g., Brooks, 1978; Jacoby & Brooks, 1984).

Similarity also plays a crucial role in models that represent categories as prototypes or as central values of their instances. Simple versions of such models assign an instance to a category if the instance meets some criterial level of similarity to the prototype or, at least, is more similar to this prototype than to those of other categories. Within the framework of these models, it seems natural to think of the similarity between instance and prototype as the *typicality* of the instance with respect to the category: The more similar an instance is to the prototype, the more typical it is of the category, and the more likely it will be classified as a category member. This relationship among similarity, typicality, and categorization makes a tidy package, since it allows us to explain in a unified way a large portion of the data from experiments on natural concepts. Rosch and others demonstrated that subjects are more likely to produce typical members as examples of a given category, to learn them earlier, and to classify them faster than atypical members. Under the simple prototype theory, these effects, and others like them, are all consequences of the similarity between the instance and the category prototype. (Smith & Medin, 1981, is the classic review of these findings.)

However, Rosch's own ideas about categories were more complex. She believed, in fact, that a prototype was merely "a convenient grammatical fiction" (Rosch, 1978, p. 40), except in the case of certain artificially generated categories. The relative typicality of an instance,

on her account, could be the result of a variety of structural principles, of which the most important is probably family resemblance of category members (Rosch & Mervis, 1975; Rosch, Simpson, & Miller, 1976). But as the term *family resemblance* implies, typicality and category membership are closely related to similarity, even in this more complicated theory. Rosch held that prototypicality is equivalent to degree of category membership; and, for any given instance, prototypicality is a function of how similar the instance is to other category members and how dissimilar it is to members of contrast categories. (Tversky, 1977, also gives an account in which the family resemblance of an instance within a category reduces to the combined similarity between that instance and other category members.)[3]

Exemplar and prototype models are tied to similarity willy-nilly; but similarity also shows up in other sorts of categorization theories in a less pure form. For instance, the *dual models* in Figure 1.1 distinguish between identification procedures that use similarity to make relatively fast, error-prone decisions and a core system that has access to deeper conceptual properties. For example, according to the feature comparison model (Smith, Shoben, & Rips, 1974), categories were supposed to be represented as sets of semantic features or attributes. We explained categorization as a two-part process, where the first part was devoted to determining the overall similarity between instance and category by comparing all of the features associated with them. The second part of this model was supposed to be more analytic, but the first part clearly committed us to the view that similarity is an important component in category decisions. What we said at the time was that the "contrast between early holistic and later analytic processing accords well with our introspections that decisions about logical matters are sometimes made quickly on a basis of similarity, while at other times decisions are the result of a more deliberative process" (Smith et al., 1974, p. 223).

Finally, at the far end of the continuum, we have network models (e.g., Collins & Loftus, 1975), frame models (Brachman & Schmolze, 1985), and Bayesian models (McCloskey & Glucksberg, 1979). All of these proposals can presumably accommodate effects of similarity on category decisions (perhaps as a by-product of other processes), but they do not give similarity a privileged role. These models are adequately described elsewhere (e.g., Rumelhart & Norman, 1988; Smith & Medin, 1981), and I will have nothing to say about them in what follows. However, their presence at the bottom of Figure 1.1 is a reminder that there are approaches to categorization other than those

based on resemblance. For the rest of this discussion, then, I'll concentrate on what we can call *pure resemblance theories*, such as exemplar and prototype models, with just a word about dual models at the end.

In sum, similarity plays a key role in many of the best-known theories of concepts, and it does so in two interrelated ways: (*a*) Similarity determines the typicality of an instance with respect to a category; and (*b*) similarity determines the probability that people will classify an instance as a category member. One way to put these ideas together is to assume first that similarity accounts for typicality; that is, typicality just *is* either similarity of an instance to a prototype or average similarity of the instance to known category members. Second, typicality in turn measures the degree of category membership. And, finally, degree of membership explains categorization probability. A trout is generally similar in size, shape, and other characteristics to other fish, and it is generally dissimilar to members of contrast classes like mammals and reptiles. Subjects therefore believe trouts to be typical fish, assume that trouts enjoy a high degree of membership in the fish category, and are very likely to categorize trouts as fish. That's the resemblance theory of concepts in a nutshell. If we can show that it is wrong, then the pure resemblance models are in trouble.

Is similarity all there is to categorizing?

In its pure form, resemblance theory claims that similarity assessment is really all there is to categorizing. The probability that a subject will assign an instance to a category is solely a function of similarity, where similarity can be computed as resemblance to one or more previously classified category members, to a prototype, or to some other representative of the category. We can also allow this similarity computation to include degree of dissimilarity to other categories. However, no other factors (apart from random error) influence category decisions. In order to be sporting, we can further concede that resemblance theory applies only to categories of natural kinds and artifacts. Clearly, resemblance theory doesn't have a chance with what Barsalou (1985; this volume) calls "goal-derived" categories such as things-to-take-with-you-on-a-vacation, since there is no reason to think that bathing suits and toothbrushes belong to this category because of their similarity to each other or to other category members. Indeed, Barsalou (1985) found that, for goal-derived categories, average similarity to other category members was not a very good predictor of an instance's rated goodness of membership within the category.

Effects of variability on categorization

Obviously, one way to disconfirm the resemblance theory is to find a factor that affects categorization but not similarity. Let me suggest that one such factor might be constraints on variability among category members. In order to clarify what I mean, consider this problem: Suppose there are two categories, one of which is relatively variable and the other relatively fixed on some physical dimension. An example might be the categories of pizzas and U.S. quarters (i.e., 25-cent pieces), since pizzas are relatively variable and quarters relatively fixed in their diameters. Now suppose there is an object about which you know only that it has a 3-inch diameter, and consider two questions about it: First, is it more likely to be a pizza or a quarter? And second, is it more similar to pizzas or to quarters? The answer to the first question clearly seems to be that the object is a pizza; for even though 3 inches is much smaller than the pizzas you usually encounter it is easy to imagine making one this small. It is harder (though not impossible) to imagine how a 3-inch quarter could come about. But, now, what about the similarity question? This one is not as clear-cut, perhaps. Yet it seems in this case that limitations governing the size of quarters do not play such a crucial role. Instead, it is plausible to take into account the simple difference in size between the 3-inch object and normal quarters or pizzas, and this may lead you to say that the object is more similar to quarters. After all, it is probably closer to the diameter of an average quarter than to the diameter of an average pizza.

To see if there really is a difference in the answers to categorization and similarity questions, we ran an experiment using 36 problems similar to the one I just mentioned. We told subjects that we were about to ask them some questions about pairs of common categories and about dimensions along which these categories varied. In every case, one of these categories was relatively fixed on the dimension in question, either by official decree or by convention, whereas the other category was relatively free to vary. For example, one trial concerned the pizzas-and-quarters pair. The diameter of quarters is presumably fixed by law, but the diameter of pizzas certainly varies widely. Other examples of categories and dimensions include: the volumes of tennis balls and teapots, the number of members in the U.S. Senate and in rock groups, the heights of volleyball nets and automobiles, and the durations of basketball games and dinner parties. The first member in each of these pairs seems relatively fixed and the second relatively variable.

We also manipulated which category – the fixed or the variable one – had normally larger values on the specified dimensions. On half of the trials, the fixed category was smaller, as in the pizza–quarter example. On the other trials, we chose the categories so that the fixed category was larger. For instance, we also included a trial concerning the diameters of basketball hoops and grapefruit, in which the fixed category (basketball hoops) has the larger values.

During an individual trial the subject's first task was to estimate the value of the largest member of the small category and the smallest member of the large category that he or she could remember. In the pizza–quarter case, for example, the subject gave us the diameter of the smallest pizza and the diameter of the largest quarter. We then told the subject that we were concerned with an object with a specific value, which we had calculated to be exactly half-way between the two extreme values he or she had named. So if the subject had said that the smallest remembered pizza was 5 inches in diameter and that the largest quarter was 1 inch, we told the subject that we were thinking of an object with a diameter of 3 inches. This intermediate value was calculated separately for each subject.

In one condition, we then asked subjects to decide which of the two categories the intermediate object belonged to; in a second condition, they chose which category the intermediate object was more typical of; and in a third condition, they chose which category the object was more similar to. For pizzas and quarters, the subject had to decide for the hypothetical object whether it was a quarter or a pizza, whether it was more typical of a quarter or a pizza, or whether it was more similar to a quarter or a pizza. This task was a between-subjects variable in this experiment, so that a given subject made only category decisions, only typicality decisions, or only similarity decisions. It is important to realize that all subjects received the same information about the categories. Moreover, there was nothing in the descriptions or instructions that encouraged subjects in the Similarity group to use one sort of property as the basis of their decision and subjects in the Typicality or Categorization groups to use some other property (except, perhaps, the very words *similarity*, *typicality*, and *category membership*).

Figure 1.2 displays subjects' choices in the three tasks, plotted as the percentage of subjects who chose the fixed category. It shows as separate functions those trials in which the fixed category was smaller (as in the pizza–quarter example) and those in which the fixed category was larger (as in the case of the basketball hoop–grapefruit pair). The first thing to notice about the results is that most subjects in the

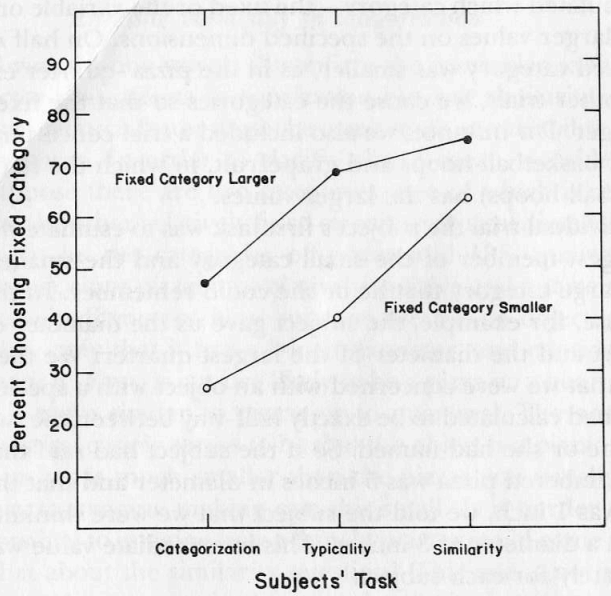

Figure 1.2. Percentage of subjects choosing the fixed category over the variable category in the Categorization, Typicality, and Similarity groups. The top function represents pairs in which the fixed category was the larger member; the bottom function represents pairs in which the fixed category was smaller.

Categorization group say that the mystery object belongs to the variable category, whereas most subjects in the Similarity group say that the object is more similar to the fixed category. The Typicality group fell in between. Over all stimulus items, 37% of the Categorization group, 54% of the Typicality group, and 69% of the Similarity group chose the fixed category. This difference is reliable, when either subjects or stimuli serve as the random variable. (This is true as well for all differences that I'll refer to as significant.)

We would expect Categorization subjects to choose the variable category if they were taking into account known constraints on the variability of these items. It may be less clear, however, why the Similarity subjects prefer the fixed category. But recall that we picked the value of the mystery item to be midway between the smallest member of the large category and the largest member of the small category. Since one of these categories is highly variable and the other fixed, the mystery instance would tend to be closer to the mean of the fixed category than to the mean of the variable category. So if the Similarity subjects were making their decision according

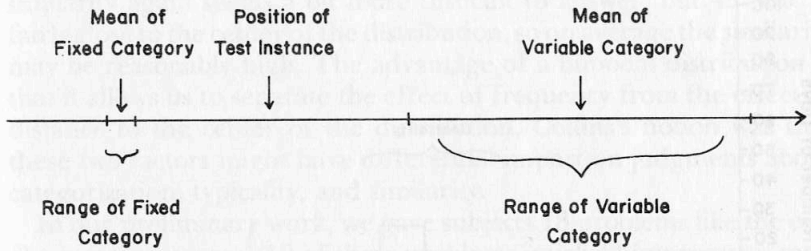

Figure 1.3. Hypothetical arrangement of fixed category, variable category, and test instance on a physical continuum.

to absolute distance between the instance and the average category value, this setup would favor the fixed category. Figure 1.3 illustrates this point. Suppose that a subject tells us that the smallest pizza is 5 inches and the largest quarter 1 inch. Then the mystery instance would be 3 inches in diameter, as mentioned earlier. Assuming that the average diameter of a pizza is 12 inches and the average diameter of a quarter is 1 inch, then the instance would be 2 inches from the average quarter but 9 inches from the average pizza.

One further fact about these data may be of interest. The distance between the two functions in Figure 1.2 shows that in all three groups subjects were biased toward choosing the larger of the two categories. Subjects were more likely to pick the fixed category given a choice between, say, basketball hoops and grapefruits than when they were given the choice between quarters and pizzas. This may be due merely to the particular categories we used as stimuli, since different category pairs appeared in these two conditions. Perhaps you could also explain this difference in terms of a Weber–Fechner function. For although the mystery item was numerically midway between the subjects' extreme values, it may have been subjectively closer to the larger category. This would have increased the chances that subjects would pick the larger category, which is the result that we obtained. A final possibility is that the difference has to do with how easy it is to imagine altering the fixedness of the fixed category. It may be intuitively easier for a fixed category to become smaller than to become larger; for example, several of the subjects in the Categorization group, when asked whether an intermediate-sized object was more likely to be a paper clip or a foot ruler, said that it was more likely to be a "broken ruler." In Kahneman and Tversky's (1982) terms, a change from large to small is a "downhill" change, whereas a change in the

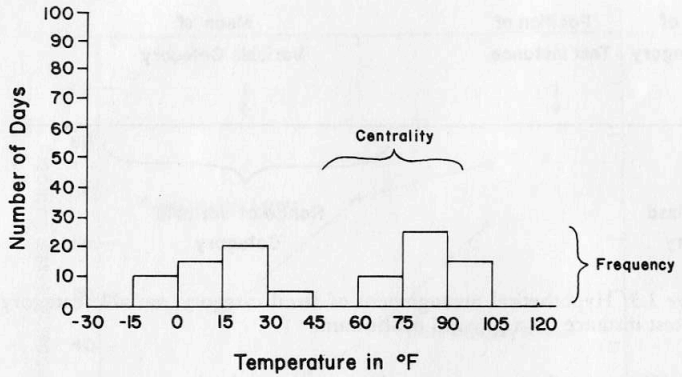

Figure 1.4. Sample histogram of the subjective distribution of daily high Chicago temperatures in January and July.

opposite direction is "uphill." If this factor enters into all three of the subjects' tasks, it could also account for the larger–smaller difference.[4]

Effects of other distributional properties

The results of the experiment just described suggest that the potential variability of category members influences categorization but has a weak impact, if any, on similarity judgments. If this is right, then it is going to be difficult to reduce categorization to similarity. Allan Collins has come up with a way to explore this relationship further, using the form of the category's density function, and it is worthwhile to report some data that Collins and I have collected (Collins & Rips, in preparation) because they provide evidence that reinforces some of the conclusions from the previous study.

To see how Collins's idea works, imagine that a meteorologist is studying daily high temperatures in Chicago, half of which occurred during the month of January and half during July. If we were to graph these temperatures, we would presumably get a bimodal distribution of values, such as the distribution in Figure 1.4. Now consider temperature readings between 45° and 60°, and ask yourself how likely it is that temperatures in this interval are among the values the meteorologist is studying, how typical they are of these temperatures, and how similar they are to the temperatures. Intuitively, the probability seems relatively low that 45°– 60° temperatures are in this set, since they fall between the peaks of the distribution. And for the same reason they may not be especially typical. The question about

similarity again seems a bit more difficult to answer; but 45°–60° is fairly close to the center of the distribution, so on average the similarity may be reasonably high. The advantage of a bimodal distribution is that it allows us to separate the effect of frequency from the effect of distance to the center of the distribution. Collins's notion was that these two factors might have differential impact on judgments about categorization, typicality, and similarity.

In our preliminary work, we gave subjects 18 problems like the one about temperatures. All of these problems involved impromptu categories composed of a mixture of elements, a mixture that we hoped would convey to subjects a bimodal distribution along a particular dimension. In addition to the temperature example, we used a problem about the weights of 100 children, half of whom were 5-year-olds and half 15-year-olds. Another problem concerned hair length of 100 teenagers, half of whom were boys and half girls.

The experiment itself consisted of two sessions. In the first one, subjects rated the likelihood, typicality, or similarity of each of a set of intervals with respect to the categories defined by the problems. On one of the trials, for example, we told our Similarity subjects that a meteorologist was studying 100 daily high temperatures in Chicago, half of which were in January and half in July. We then asked them to rate the similarity of temperatures within particular intervals to the temperatures in this set. They rated the similarity of temperatures between −30° and −15° to the temperatures studied by the meteorologist, the similarity of temperatures between −15° and 0°, between 0° and 15°, and so on. There were 10 to 11 such intervals for each problem, spanning what we hoped would be the relevant range of values (see Figure 1.4). Subjects in the Typicality group received problems of the same sort, but they rated the typicality of each interval with respect to the set. Finally, subjects in the Categorization group rated the likelihood that temperatures in each interval were among the temperatures in the set. Each group contained 12 subjects. In the second session of the experiment, subjects from all three groups received a description of the same categories in a new random order. But this time we gave them graph paper and asked them to draw a histogram for these categories. For instance, one of their problems was to draw a histogram of the distribution of temperatures that the meteorologist was studying. The intervals that we marked off on the base of the histogram corresponded to the intervals they had been quizzed about in the earlier session.

For each problem and for each subject, then, we have two kinds of information: a frequency histogram and ratings of similarity, typi-

Table 1.1. *Mean standardized regression weights (β's) for ratings as a function of histogram frequency and centrality*

	Rating type		
	Categorization	Typicality	Similarity
Frequency	.47	.41	.33
Centrality	−.01	.09	.17

cality, or category likelihood. Our aim was to see whether aspects of the distribution differentially influenced the ratings, and we therefore extracted two measures from each of the distributions. One measure was simply the height of the histogram at each interval; the other was a measure of how close that interval was to the median value of the histogram.[5] I'll call the first of these measures the *frequency* of the interval and the second the *centrality* of the interval. Figure 1.4 illustrates these measures for a sample histogram of the temperature problem.

Our next step was to perform regression analyses in which the ratings served as the dependent variable and the frequency and centrality measures served as independent variables. We carried out three separate regressions, one each for the Similarity, Typicality, and Categorization groups. Table 1.1 lists the β weights (i.e., standardized regression coefficients) from these analyses. It is apparent that the categorization judgments are quite sensitive to frequency; so if an interval is near one of the peaks of the histogram, subjects tend to rate the interval as likely to be in the category. On the other hand, categorization ratings do not depend on how close the value is to the histogram's center. A value is just as likely (or unlikely) to be a category member whether it is in the middle of the distribution or is at an equally frequent point in the tails. For typicality judgments, there is a hint of a centrality effect and slightly less dependence on frequency. The similarity ratings continue this trend with a more robust centrality effect and a weakened effect of frequency. The interaction between rating type and measure (frequency or centrality) is significant in these data.

Implications

The two studies I have just described have some common properties that are worth noticing. In the first place, both experiments

suggest that categorization is sensitive to distributional properties in a way that similarity decisions are not. In Study 1 this was manifest in the Categorization group's choice of the variable over the fixed category and in Study 2 by the correlation between frequency and category ratings. This evidence is also consistent with earlier experiments by Fried and Holyoak (1984), using artificial categories consisting of grids of filled and unfilled cells. They found, in particular, that subjects tended to classify test instances as members of a high-variable rather than a low-variable category, even when the instances were physically closer to the low-variable category's prototype. If category decisions, but not similarity decisions, depend on variability and like factors, then categorization cannot be equated with similarity.

A second commonality between the experiments is that similarity responses appear to depend on distance to the categories' central values. This relationship explains why the Similarity group in the first study preferred the fixed category to the variable one. It showed up again in the second study as a correlation between similarity and centrality. As we noted in the first part of this discussion, resemblance theories propose that people categorize instances in terms of distance to the categories' central tendency. It is ironic, then, that the results of the second study suggest that although centrality affects similarity judgments, it has no effect on categorization. This centrality factor also hints that we may be able to alter the similarity of an instance to a category without changing the probability that subjects will classify it as a category member. Experiments that I describe later pursue this hint.

A third common feature of these experiments is that typicality decisions appear to be a compromise between categorization and similarity. The usual story is that similarity is responsible for variations in typicality and that typicality itself is simply a measure of degree of category membership. However, the results so far suggest that although typicality may share properties with both similarity and probability of category membership it is identical with neither. As we shall see, this last conclusion will have to be reassessed in light of later results. Nevertheless, these parallels between the experiments are reassuring. Because you might object to the unnaturalness of forced choice in the first study, it is helpful to find similar results for ratings in the second. Likewise, the somewhat artificial mixture of categories in the second study is balanced by the more ordinary categories in the first.

Of course, we still need to be cautious before settling on an

interpretation of these studies. One point that bears emphasis is that these data do not imply that similarity plays *no* part in category decisions. It is certainly possible that subjects' similarity judgments and their category judgments both depend on some common process. Maybe you could even call this a *similarity* or *resemblance* computation without doing too much violence to these terms. Likewise, similarity could sometimes serve as a heuristic for category assignment (as Medin & Ortony suggest in chapter 6). Neither possibility is at stake here. What I believe the studies do show, however, is that category decisions are not *solely* a matter of similarity (even in the special sense of a common underlying process): If they were, then factors like variability that affect categorization should also affect similarity judgments. Because resemblance theory is precisely the claim that categorization can be reduced to similarity alone, resemblance theory must be false.[6]

However, there is one way of salvaging resemblance theory that we need to consider carefully. This is the idea that resemblance itself may come in several varieties, one of which is responsible for categorization judgments and another for similarity judgments. For example, one explanation of the first study is that subjects in both the Categorization and the Similarity groups were computing the similarity between the mystery instance and the means (or other central values) of the two categories. However, Categorization subjects, according to this hypothesis, used normed distance, with distance to the variable category normed by the variable category's standard deviation and distance to the fixed category normed by the fixed category's standard deviation. Similarity subjects, on the other hand, used unnormed distance to the two means.

But, although this hypothesis accounts for the obtained difference, it does not accord with subjects' own view of the matter. In the first study, we asked subjects to talk aloud as they made their choice, and we recorded their responses. When Categorization subjects chose the variable category, they typically said that the mystery object *must* be a member of that category because members of the fixed category *can't* be that size. Here are some examples:

Subject A: Is something with 170 members an English alphabet or a bowl of rice? "An English alphabet is restricted to 26 letters, but a bowl of rice can be any size; so it must be a bowl of rice."

Subject B: Is an object that holds 1.75 cups a teapot or a tennis ball? "I'd say a teapot – a smaller teapot – because tennis balls would have to be the same size."

Is someone 20 years old a master chef or a cub scout? "I'd say he'd be a master chef, because I mean cub scouts – I mean the things like nickels [in an earlier problem] and things like that just seem real – I mean they just *can't* be outside of that age group or they're not even that thing anymore. So I can imagine a chef that's 20 years old, but it's just hard to imagine a cub scout that's that old."

Subject C: Is something 4.75 feet high a stop sign or a cereal box? "It would probably be one huge cereal box, because no one would see the stop sign if it were that small. . . . A stop sign would have to be a certain height, and while I wouldn't expect to see a cereal box that big, it wouldn't make sense to have such a small sign."

Subject D: Is something 18.75 hours long Valentine's day or a final exam? "A final exam. I was thinking about that one and it wouldn't be a day or a Valentine's day according to any definition if it was 18 hours; so it would have to be some sadistic final exam."

Is something with 47 members a jar of pickles or a deck of playing cards? "A jar of pickles, because if a deck of playing cards didn't have 52 or greater cards it wouldn't *be* a deck of playing cards."

Subjects' modal constructs in these examples – "it *must* be a bowl of rice"; "they just *can't* be outside of that age group or *they're not even that thing*"; "*would have to be* a certain height"; "it *wouldn't* be a day *according to any definition*" – indicate that they are engaging in a form of reasoning that goes beyond simple distance comparison. Subject C, for example, does not justify his choice on the grounds that 4.75 feet is closer to the size of a cereal box than to the size of a stop sign. Instead, he believes that something of that height simply can't be a stop sign ("it wouldn't make sense"); hence, by elimination, it must be a cereal box, which doesn't have these limitations on height. In other words, the relevant restrictions don't appear to be altering the subjects' sense of what's similar to what, but instead they act directly to rule out the very possibility that the mystery instance is a category member. For Subject C the restriction is a functional one (people would have trouble seeing such a short stop sign); in other cases it appears to be definitional (a day is defined as 24 hours long) or conventional (tennis balls have to be the same size).

Of course, it may well be true that there are multiple types of similarity that come into play at different stages of development (L. Smith, this volume) or for different purposes (Gentner, this volume); however, there is little evidence that such differences account for the

above results. We will return to the multiple-similarity idea at the end of this chapter, armed with evidence from some additional studies.

The independence of similarity and category judgments

> *The Callitrices differ from the rest [of the monkeys] in nearly their whole appearance.*
>
> From a Latin bestiary (trans. White, 1954)

I have been arguing that categorizing may be more complex than resemblance theory would lead you to believe. In retrospect, it is not surprising that the purest of the pure resemblance models have been developed to account for classifying artificial categories of dot patterns, schematic faces, letter strings, and the like; subjects in these experiments have no information about the categories except what they can extract during the learning trials. But, for categories like eggs or quarters or Chicago temperatures, subjects know a lot; and some of this knowledge may simply be too abstract or too extrinsic to contribute to the categories' similarity, at least according to the meaning that subjects attach to *similarity*.

This way of viewing the issue suggests that similarity and categorization may be independent relations, in the sense that we can manipulate one of them with only minimal effects on the other. To investigate this question, we have looked for situations in which similarity could change without changing classification. And to provide a contrast we have also tried to construct cases in which an instance switches categories without a change in similarity. The first type of situation is one where an object's properties are accidentally transformed in such a way that they begin to resemble those of members of a different category. The second situation occurs when an object's properties are altered in a more essential way. Of course, *essential* in this context does not mean that the properties are necessarily true (*de re*) of the instance itself. An essential change is simply one that subjects believe is important to the instance's membership in a specified category, and an accidental change is one that is not important in this way.[7] For the sake of generality, we have carried out two experiments of this sort, one with natural kinds and the other with categories of artifacts.

Transformations on natural kinds

We tried to simulate the relevant situations for our subjects by means of a group of stories describing transformations that happen to imaginary animals. In the *Accident* condition, each story described a hypothetical animal in such a way that subjects were likely to identify it as a member of a particular category – either birds, fish, insects, reptiles, or mammals. The story then described the animal as undergoing some catastrophe that caused many of its surface properties to resemble those of one of the other categories. For example, an animal that started out as a bird might come to have some insect properties. The actual category labels *bird* and *insect*, however, did not appear in the story. The subjects were then asked to rate whether the animal was more likely to be a bird or an insect, whether it was more typical of a bird or an insect, and whether it was more similar to a bird or an insect. The subjects circled a number on a 10-point rating scale that had *bird* on one end and *insect* on the other. Each subject read and rated 10 stories, with each story corresponding to a different pair drawn from the five categories. Here, for example, is the story of the bird who became insectlike:

> There was an animal called a sorp which, when it was fully grown, was like other sorps, having a diet which consisted of seeds and berries found on the ground or on plants. The sorp had two wings, two legs, and lived in a nest high in the branches of a tree. Its nest was composed of twigs and other fibrous plant material. This sorp was covered with bluish-gray feathers.
>
> The sorp's nest happened to be not too far from a place where hazardous chemicals were buried. The chemicals contaminated the vegetation that the sorp ate, and as time went by it gradually began to change. The sorp shed its feathers and sprouted a new set of wings composed of a transparent membrane. The sorp abandoned its nest, developed a brittle iridescent outer shell, and grew two more pairs of legs. At the tip of each of the sorp's six legs an adhesive pad was formed so that it was able to hold onto smooth surfaces; for example, the sorp learned to take shelter during rainstorms by clinging upside down to the undersides of tree leaves. The sorp eventually sustained itself entirely on the nectar of flowers.
>
> Eventually this sorp mated with a normal female sorp one spring. The female laid the fertilized eggs in her nest and incubated them for three weeks. After that time normal young sorps broke out of their shells.

We hoped that this story – besides alerting our subjects to the hazards of toxic wastes – would get them to rate the sorp more likely to be a bird but more similar to an insect. Half of the subjects in the Accident condition read the story just mentioned in which the sorp begins life with bird properties and ends with some insect properties. The other subjects read a similar story about an animal that begins with insect properties and acquires bird properties. For purposes of comparison, an Accident control group read a shortened form of the same stories, which described only the first, precatastrophe part of the animals' lives. These subjects, we felt sure, would rate the animal as uniformly

more likely to be, more typical of, and more similar to the initial category. For example, we gave these control subjects the first, birdy part of the sorp description and expected them to rate the sorp as more likely to be a bird, more typical of a bird, and more similar to a bird.

In addition to the Accident condition and its control, we also included an *Essence* condition, whose purpose was to influence categorization ratings without influencing similarity. Subjects in this condition also read stories about animals that undergo some radical transformation; but this time the change was the result of maturation, similar to the change from a tadpole to a frog or from a caterpillar to a butterfly. The two halves of an animal's life were given separate names in order to mark this distinction. For example, this is the sorp's story as it appeared in the Essence condition:

> During an early stage of the doon's life it is known as a sorp. A sorp's diet mainly consists of seeds and berries found on the ground or on plants. A sorp has two wings, two legs, and lives in a nest high in the branches of a tree. Its nest is composed of twigs and other fibrous plant material. A sorp is covered with bluish-gray feathers.
>
> After a few months, the doon sheds its feathers, revealing that its wings are composed of a transparent membrane. The doon abandons its nest, develops a brittle, iridescent outer shell, and grows two more pairs of legs. At the tip of each of the doon's six legs an adhesive pad is formed so that it can hold onto smooth surfaces; for example, doons take shelter during rainstorms by clinging upside down to the undersides of tree leaves. A doon sustains itself entirely on the nectar of flowers.
>
> Doons mate in the late summer. The female doon deposits the eggs among thick vegetation where they will remain in safety until they hatch.

After reading each story, our subjects rated the similarity, typicality, and likely category membership of the animal's first stage. After the sorp/doon story, the subjects rated whether the sorp – not the doon – was more similar to a bird or an insect, more typical of a bird or an insect, and more likely to be a bird or an insect. We also included an Essence control condition, in which a new group of subjects read descriptions of only the animals' early stage and rated the same three dimensions. Each group contained 24 subjects. An individual subject always rated the dimensions in the same order; however, within a group, four subjects were assigned to each of the six permutations.

The results of this experiment are easy to relate. Let's look first at the mean ratings from the Accident condition and its control in Figure 1.5. The scale on the *y* axis is oriented in the figure so that high ratings correspond to the category of the animals' original appearance. For example, in the case of the sorp, which started off as a bird and

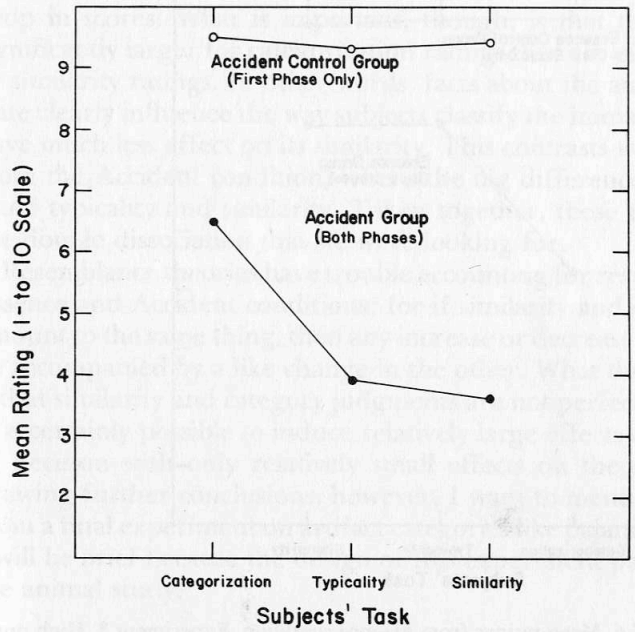

Figure 1.5. Mean ratings from Accident condition, Experiment 3. High numbers represent the category most like the animals' original appearance; low numbers, the category most like their later appearance. Top function is from Accident control subjects, and bottom function from Accident experimental subjects.

developed insect properties, high numbers indicate that subjects rated it as a bird, whereas low numbers indicate that they rated it as an insect. The control group (open circles in the figure) provides a base line in this experiment. This group read only the first-phase descriptions; and, as expected, they thought the animals similar to, typical of, and more likely to be members of this first-phase category. However, the experimental group's knowledge of the animals' later mishaps caused the ratings to shift in the direction of the alternative category, with a much greater drop for typicality and similarity ratings than for categorization ratings. Although we were unable to change similarity ratings without changing categorization at all, the interaction between groups and rating tasks is nevertheless reliable in these data. Notice, in particular, that the experimental subjects tended to think that a transformed animal is more likely to be a member of its first-phase category, but more similar to and more typical of the category whose properties it develops.

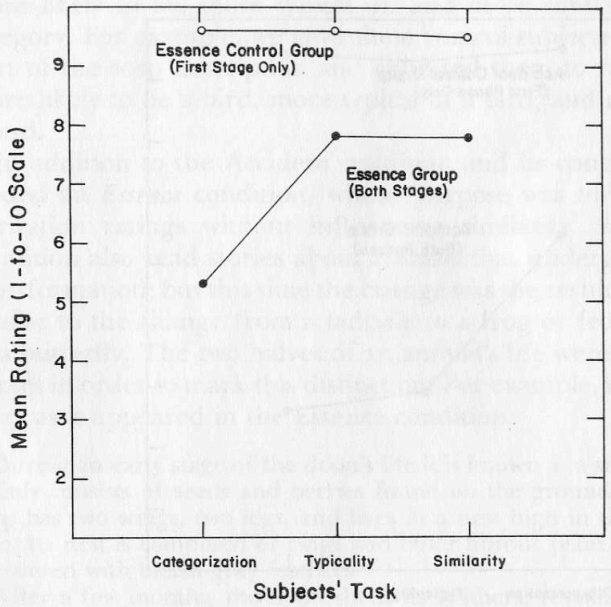

Figure 1.6. Mean ratings from Essence condition, Experiment 3. High numbers represent the category most like the animals' original appearance; low numbers, the category most like their later appearance. Top function is from Essence control subjects, and bottom function from Essence experimental subjects.

It is also worth noting that in this experiment (as well as in later ones) mean typicality is nearly equal to the similarity ratings. This differs from the results of the previous experiments in which typicality was somewhere between similarity and categorization. The change may be the result of moving from a between-subjects design to a within-subjects design on the three types of ratings. In none of the experiments, however, is typicality equivalent to category ratings, contrary to what you might expect on the assumption that typicality measures degree of category membership.

When we turn to the data from the Essence condition, we find a very different pattern of results. Figure 1.6 plots the mean ratings so that high numbers again correspond to the category most like the animals' initial description. For the birdlike sorp whose metamorphosis changed it into an insectlike doon, high numbers mean that subjects rated it as a bird and low numbers that they rated it as an insect. As before, the control group knew only about the first-stage properties, and their ratings are at ceiling. The experimental subjects

learned about the animals' transformation, and this caused an obvious drop in scores. What is important, though, is that the decrease is significantly larger for categorization ratings than for either typicality or similarity ratings. In other words, facts about the animal's mature state clearly influence the way subjects classify the immature form but have much less effect on its similarity. This contrasts with the results from the Accident condition, where the big difference appeared in rated typicality and similarity. Taken together, these results give us the double dissociation that we were looking for.

Resemblance theories have trouble accounting for results from both Essence and Accident conditions; for if similarity and categorization amount to the same thing, then any increase or decrease in one should be accompanied by a like change in the other. What the results show is that similarity and category judgments are not perfectly correlated. It is certainly possible to induce relatively large effects on either type of decision with only relatively small effects on the other. Before drawing further conclusions, however, I want to mention the results from a final experiment on artifact categories like pajamas and radios. I will be brief because the design of this experiment parallels that of the animal study.

Transformations of artifacts

The distinction between essential and accidental changes seems more clear-cut in the case of natural kinds than in the case of artifacts. Natural kinds like reptiles or sugar or quasars have inner natures that can support lawlike generalizations (e.g., All reptiles are cold-blooded) and counterfactual conditionals (e.g., If Rudolph were a reptile, he'd be cold-blooded). Artifacts, on the other hand, do not have inner natures, at least on some theories of these objects (Schwartz, 1980). It is certainly unlikely that scientists would ever seriously study the nature of, say, umbrellas; and likewise, *umbrella* is hardly the kind of term that we would expect to show up in scientific laws. If subjects adopt this view of artifacts, they may conceive of all changes to these objects as accidental changes; they may believe that there simply isn't any essence to change. On the other hand, certain properties of artifacts, although perhaps not strictly necessary, are clearly important in qualifying an object as a category member. It is possible that changes to these properties may be enough to shake subjects' confidence in the object's membership status. In the protocol examples from the first study, Subject D apparently takes this attitude toward decks of playing cards, saying that collections with less than 52 cards "wouldn't

be a deck of playing cards." Changes to properties of this sort may produce the same effects as essential changes to natural kinds.

Of course, accidental changes are easy to produce, since artifacts can undergo all sorts of surface alterations without necessarily becoming a member of a new category. To cite one of the stimulus stories, you can imagine altering an umbrella so that it looks much like a lampshade. As long as the object is still used to keep off the rain, it remains an umbrella; it has not switched categories. The actual story read like this:

> Carol Q. has an object which is a collapsible fabric dome. It consists of multicolored waterproof fabric stretched taut across six metal struts radiating from a central post in the dome. The metal struts are jointed so that they may be folded and this allows the fabric dome to be collapsed. When fully extended the dome is about three feet wide. Carol uses this object to protect her from getting wet when she is walking in the rain.
>
> Carol saw an article in a fashion magazine about a new style for objects such as this which she copied with her own. She added a pale pink satin covering to the outside surface, gathering it at the top and at the bottom so that it has pleats. Around the bottom edge of the object she attached a satin fringe. To the inside of the dome at the top she attached a circular frame that at its center holds a light bulb. Carol still uses this object to protect her from the rain.

As expected, it proved more difficult to come up with essential changes in an artifact while preserving its similarity to its original category. Our first thought was that we could do this by stipulating a new function for the object – for example, by describing an umbrella that someone comes to use as a lampshade. But this ploy didn't work. Subjects in a pilot study insisted that the umbrella remained an umbrella no matter how people happened to use it. Clearly, function is not the sole criterion for classifying artifacts. Eventually, it occurred to us that a better way to produce an essential change in an artifact was to specify the intentions of the designer who produced it. For example, we could describe an object that looks exactly like an umbrella; then, by telling subjects that its designer meant it as a lampshade we might be able to convince them that it was a lampshade that just happens to resemble an umbrella. Here is the umbrella/lampshade story we constructed:

> Carol Q. of CMR Manufacturing designed an object which is a collapsible fabric dome. It consists of multicolored waterproof fabric stretched taut across six metal struts radiating from a central post in the dome. The metal struts are jointed so that they may be folded and this allows the fabric dome to be collapsed. When fully extended the dome is about three feet wide.
>
> Carol intended for this object to be used with the inside of the dome facing up as an attachment to ceiling light fixtures. Attached in that way the multicolored fabric filters the light emanating from an overhead light fixture.

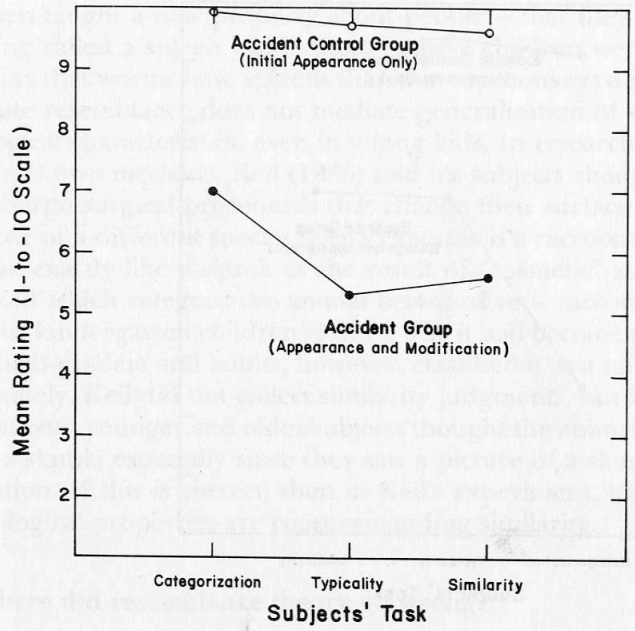

Figure 1.7. Mean ratings from accident condition, Experiment 4. High numbers represent the category most like the artifacts' original description; low numbers, the category most like their later appearance. Top function is from Accident control subjects, and bottom function from Accident experimental subjects.

We composed 18 stories describing accidental changes, each story dealing with a different pair of common artifacts. An additional 18 stories described essential changes to the same artifact pairs. As in the previous experiment, these two sets of stories were given to separate groups of subjects, who rated similarity, typicality, and likelihood of category membership. We also tested two control groups, who received just the first parts of the object descriptions. Each of the four groups contained 12 subjects.

Mean ratings from the Accident group and its control appear in Figure 1.7, with larger values on the *y* axis again indicating ratings congruent with the instance's original category. Although the effect is not as dramatic as it was for the natural categories, we were able to decrease similarity and typicality ratings significantly more than category ratings. Carol's fashionable umbrella – decorated with satin

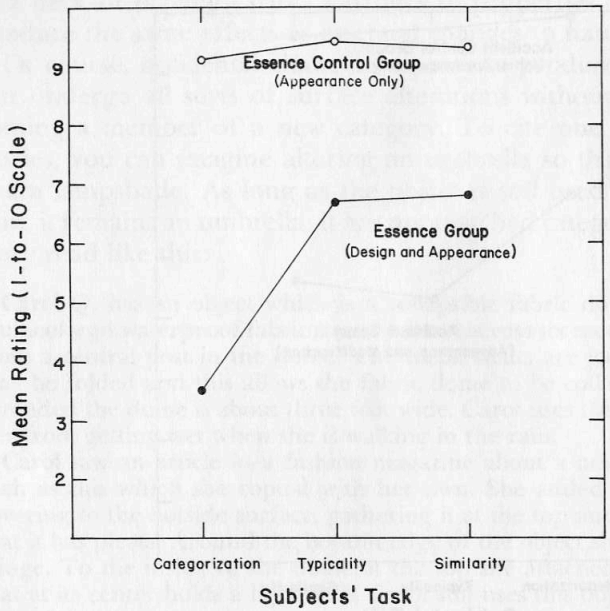

Figure 1.8. Mean ratings from Essence condition, Experiment 4. High numbers represent the category most like the artifacts' original description; low numbers, the category most like the designers' intention. Top function is from Essence control subjects, and bottom function from Essence experimental subjects.

ruffles and a light bulb – is still an umbrella but no longer especially similar to or typical of one. We can compare these results to those in Figure 1.8 from the Essence group and its control. Specifying the intentions of the object's designer completely changed the way subjects classified the object, even though the object still resembled members of the alternative category. If Carol designs an object as a lampshade, then it *is* a lampshade despite looking much like an umbrella.

Both natural kinds and artifacts, then, exhibit some independence between properties that make them resemble members of this or that category and properties that qualify them as category members. The essential properties for natural kinds apparently have to do with the mature, reproductive state of the organism, whereas the essential properties of artifacts lie in the intentions of their designers. The results for natural kinds have some precedents in the developmental literature. For example, Carey (1982) notes that 4-year-olds rate people as being more similar to toy monkeys than to (real) worms. Yet,

when taught a new property about people – that they have a green thing called a spleen inside them – these children were more likely to say that worms have spleens than that toy monkeys do. Apparently, brute resemblance does not mediate generalization of specifically biological characteristics, even in young kids. In research that is closer to our own methods, Keil (1986) told his subjects about animals that undergo surgical procedures that change their surface properties to those of a different species. Keil's example is a raccoon that comes to look exactly like a skunk as the result of "cosmetic" surgery. When asked which category the animal belonged to – raccoon or skunk – most kindergarten children insisted that it had become a skunk; most fourth-graders and adults, however, classified it as a raccoon. Unfortunately, Keil did not collect similarity judgments, but it seems likely that both younger and older subjects thought the animal more similar to a skunk, especially since they saw a picture of a skunk as an illustration. If this is correct, then in Keil's experiment, too, underlying biological properties are countermanding similarity.

Where did resemblance theory go wrong?

Resemblance theory suffers two deficiencies. First, it has trouble with factors such as variability, frequency, underlying biological properties, and personal intentions that affect how an instance is classified but do not much affect the instance's similarity. These properties have a common characteristic: They are all hidden from a casual view of the instance. None are obvious parts of the instance, either because they are relational or because they are discernible only with scientific acumen. It is likely that this extrinsic or nonpart status disqualifies these properties (in our subjects' opinion) from contributing to the instance's similarity to another instance or category. The second difficulty is that resemblance theory cannot explain why aspects of an instance's appearance should affect the similarity of the instance to a category but not the probability that it is a category member. These surface features are part of the instance itself and, therefore, play a role in similarity computations; they simply are not important in determining the instance's category membership.

As noted earlier, none of this shows that similarity *never* plays a role in classifying things. In many ordinary settings, properties that make an instance similar to a category probably also furnish clues to its category membership. The point is that these clues are not definitive; they are only presumptive. They are not sufficient, since there are instances that are highly similar to a category without being members

(as in the case of Carol the designer's lampshade that looks like an umbrella). And they are not necessary, since there are bona fide category members that look more like members of some other species (as in the case of Keil's surgically altered raccoon or our bird who accidentally came to look like an insect). The idea that similarity is a fallible clue or heuristic for categorization is probably the inspiration for dual models (Miller & Johnson-Laird, 1976; Smith et al., 1974), which I mentioned at the beginning of this chapter. Nothing in these data refutes models of this type, as long as the two components are loosely coupled. Likewise, similarity may well provide an important heuristic in decision making (Smith & Osherson, this volume), conceptual combination (Smith, Osherson, Rips, & Keane, in press), and other higher cognitive tasks. It's *pure* resemblance models that are the casualties of the present experiments.

Objections and responses

Before giving up pure resemblance, however, we should consider what resemblance theorists might say in their own defense. There appear to be three lines of argument, which amount to objections to the evidence I've presented here. The first is that we have considered the wrong kinds of categories, and the second, that we have looked at the wrong kind of categorization task. The third objection takes up the multiple-similarities idea that we touched on at the end of the second section.

Objection 1: two kinds of categories. One line that resemblance theorists might adopt is to deny the relevance of the stimuli. They might say, for example, that resemblance theory was never intended to account for the kinds of examples that these studies employed. They could point to the sci-fi quality of the stimuli and claim that resemblance theory need not apply to these radically contrary-to-fact categories.

It is true, of course, that many of the instances that we constructed are not ones you are likely to come across. Birdlike creatures don't transform themselves into insects, and people don't adorn their umbrellas with light bulbs. Our aim in creating these examples was to drive a wedge between resemblance and categorization and not to mimic real objects. However, I think it is possible to maintain that cases like the ones we studied are not all that rare: Although birds don't change into insects, fishlike objects do turn into frogs and wormlike ones into butterflies. There are also lots of cases in which artifacts look like members of other categories. Christmas catalogs list jewelry boxes in the form of heads of lettuce, candles that look like pieces of fruit, cameras that look like cigarette packs, and so on. Resemblance theory cannot dodge all such examples without seriously weakening its own credibility.

Objection 2: two kinds of categorization. Another objection along the same lines challenges the nature of the experimental tasks. The idea is that the most representative cases of categorization are ones that occur quite rapidly, on the order of a few hundred milliseconds. You see an object and immediately recognize it as an egg. You don't stop to consider whether it might be a Russian ornament that some clever jeweler has designed. By contrast, the studies I have described obviously call for deeper reasoning or problem solving on the subjects' part. It is quite possible that similarity is responsible for immediate categorization, even if it does not suffice for these more complex situations.

It may well be correct that immediate perceptual categorization often rests on similarity. Thus, if the categorizing situation is time-limited so that a subject has access only to the surface properties of the instance, then these surface properties will dominate the decision. Since overlap on surface properties seems to be the hallmark of similarity (given our own results), it is quite natural to think of this sort of categorization as similarity-based.

Still, we need to be careful about resemblance theory even for on-the-spot perceptual categorizing. What seems to be rapid classification based on surface properties may well turn out to involve procedures more complicated than similarity matching. For example, consider the process of assigning a visually presented object to a superordinate category (e.g., furniture) or a goal-derived category (e.g., birthday presents). Assuming that people can assign instances to these categories rapidly, they must be doing so on the basis of processes that go beyond similarity. Similarity may help them decide that something is a bunch of flowers, perhaps; but they must then make additional inferences to decide that this is a possible birthday present. Along the same lines, imagine a perceptual version of the first study, in which subjects see, for example, an outlined circle representing the mystery instance and decide whether something of this size and shape is more similar to, more typical of, or more likely to be a pizza or a quarter. There is no reason to think that the results of such a study would be any different from those we have obtained.

The real danger in this approach, however, is taking fast-paced

object recognition as the archetype for categorization. Countless category decisions are not of this type. We classify people as friendly or hostile, arguments as convincing or fallacious, numbers as prime or composite, investments as safe or risky, policies as fair or biased, governments as socialist or totalitarian, cultures as advanced or primitive, religions as orthodox or heterodox, documents as authentic or forgeries, crimes as felonies or misdemeanors, diseases as chronic or acute, purchases as expensive or cheap, jokes as funny or offensive, speeches as informative or boring, jobs as rewarding or make-work, vacations as relaxing or vexing. Even if it turns out that first-glance object recognition is driven by similarity, we would need additional arguments (convincing ones) to show that this kind of classification has a privileged status that should be taken as a model for all categorizing.

Objection 3: two kinds of similarity. In discussing the first two studies, we considered the possibility that the results of the Similarity and Categorization conditions were both due to similarity but with a different metric for the two types of judgment. Of course, simple transformations of distance or similarity would not account for the results of the last two experiments, but there might be a generalization of this idea that would work. A proponent of resemblance theory might say, for example, that subjects in the Accident and Essence conditions used similarity to determine both their similarity and categorization ratings; however, for the similarity ratings they weighted surface properties especially heavily, whereas for the categorization ratings they weighted underlying properties more heavily.

It is easy to make fun of proposals of this kind, since they seem contrived to get the theory out of trouble. In other words, this objection sounds as if Resemblance theorists are trying to convince us that there are really two kinds of similarity: "categorization similarity," which is involved in classifying things, and "similarity similarity," which is involved in ordinary similarity judgments. This seems absurd, since there appears to be no justification for extending the meaning of *similarity* in this way. (It is a bit like saying that all life is a dream – but some dreams are waking dreams and some are sleeping dreams.) But perhaps we should not be too quick to dismiss this objection. I think people are tempted by this line because they believe that categorizing things and judging their similarity have some significant points in common. As discussed earlier, both processes might inspect predicates that are true of the objects in question, and both might compare these predicates in certain ways. It would be nice to have a

term to refer to these shared factors, and *similarity* and *resemblance* easily come to mind.

Certainly none of the evidence that I have presented contradicts the idea that categorizing and judging similarity have some commonalities, and those who wish to use *similarity* or *resemblance* in a technical sense to mark these commonalities are free to do so. However, there are two serious problems with this way of thinking. One problem is that this technical sense of *similarity* is nearly vacuous as a psychological explanation. Consulting and comparing mental predicates takes place, not only in categorizing and in judging similarity but also in nearly every other kind of cognitive task: language comprehension, memory search, reasoning, problem solving, decision making, and so on. Until this sense of *similarity* is made more specific, it cannot shed much light on the process of categorizing, contrary to the claims of resemblance theory. The second deficiency is that this kind of similarity may be open to the sort of circularity arguments advanced by Murphy and Medin (1985) and others. If similarity is cognitively primitive and perceptually given, then you might hope to use it in order to reconstruct categorization, without worrying that category knowledge will alter similarity itself. But if similarity is simply predicate comparison, then all bets are off. For example, if you explain why people classify bats as mammals by saying that bats are similar to other mammals, you cannot simultaneously explain that similarity by invoking shared predicates such as *is a mammal.* (This is not to say, however, that all predicate-comparison theories are circular; see n. 1.)

The implication is this: On the one hand, if similarity denotes something like raw perceptual resemblance (which is what it seems to mean to most subjects), then resemblance theory is not, by itself, powerful enough to explain categorization. That is what the four studies establish. On the other hand, if similarity merely means predicate comparison, then resemblance theory risks vacuity and circularity. There may be a way of slipping through this dilemma, but it is up to the resemblance theorist to show us how this can be done.[8]

Categorization as explanation

You can get a good glimpse of what is wrong with resemblance theory by considering an example of Murphy and Medin's. Imagine a man at a party who jumps into a swimming pool with all his clothes on. You might well classify such a person as drunk, not because he is

similar in some way to other drunk people or to a drunk prototype, but because drunkenness serves to explain his behavior. As Murphy and Medin point out, classification in this case is a matter of an inference about the causes of the action we witnessed.

But notice that this example generalizes easily to other category decisions. In most situations that call for categorizing, we confront some representation of an instance with our knowledge of the various categories it might belong to. If the assumption that the instance is in one of these categories provides a reasonable explanation of the information we have about it and if this explanation is better than that provided by other candidate categories, then we will infer that that instance is a member of the first category. For example, in deciding whether an object with a 3-inch diameter is a quarter or a pizza, we might consider alternative stories about how a pizza or a quarter of that size could come about. Because the pizza explanation is probably more parsimonious than the quarter explanation, we will infer that the object is indeed a pizza.[9]

The results of the last two experiments also yield to this sort of analysis. For example, if we learn about a birdlike animal that turns insectlike as the result of a chemical accident, then we might well consider the possibility that the chemicals modified the superficial appearance of the animal, leaving the genetic structure unchanged. Since this explanation seems a bit more plausible than the alternative possibility that the chemicals actually changed the genetic structure, we are likely to decide that the animal is a bird. On the other hand, suppose we are told about a birdlike creature that matures into one that is insectlike. Given what we know about biological development, it seems reasonable to suppose that the later stage is indicative of the animal's true category. Hence, the hypothesis that the creature is an insect may provide a better explanation of this instance than the hypothesis that it is a bird.

Of course, this way of thinking about categorizing is not very close to a true cognitive model. In order to fill in the details, we would need an account of how people generate explanations and how they evaluate them. Unfortunately, these problems have proved extremely difficult ones in philosophy of science (see, e.g., Achinstein, 1983), and there is no reason to think that they will be any easier within a psychological framework (Fodor, 1983). To make matters worse, this type of account may be open to problems of circularity in much the same way as the predicate-comparison idea, discussed in the preceding section. But despite these difficulties, an explanation-based approach to categorizing is worth taking seriously, partly because similar processes

are required for other cognitive abilities. For example, Schank, Collins, and Hunter (1986) have argued that category learning also depends on constructing explanations. In particular, mistakes in classifying objects or events should sometimes cause us to modify our beliefs about the nature of the relevant category. It is not always easy, though, to determine exactly which beliefs led to the error, and in these situations we may have to search for a plausible explanation of how the difficulty came about. In other words, category learning is often like troubleshooting a mechanical device, requiring similar explanations and tests. But if explanation figures into the way we learn about categories, it would not be surprising if it also played a role in the categorizing process itself.

Second, explanation is also needed as part of an account of many forms of nondeductive reasoning (Harman, 1966, 1986). Scientific inference, for example, is generally a matter of accepting the truth of the hypothesis that gives the best explanation of the data at hand. Indeed, categorization and category learning are special cases of inference to the best explanation. In the Murphy–Medin example, we conclude that the partygoer is drunk because drunkenness provides a good account of why he jumped into the pool. In the pizza–quarter example, we conclude that the 3-inch object must be a pizza rather than a quarter because we can more easily explain how a pizza of that size could be created. One way to see the connection between categorization and inference to the best explanation is to notice that many inference problems can be turned into categorization problems by a minor change of wording. The question whether all material of a given type conducts electricity is equivalent to asking whether it is a member of the category of electricity-conducting objects. To answer questions like this, we typically use evidence about the underlying nature of the objects and their lawful interrelations. Mere similarity is too weak to solve these inductive problems.

The idea that categorizing is a form of inference bears an obvious kinship to other recent hypotheses about the nature of concepts. Many investigators have proposed that similarity-based heuristics cannot possibly account for all uses of concepts. As a result, these investigators have postulated concept cores that figure in language understanding (Miller & Johnson-Laird, 1976) and conceptual combination (Osherson & Smith, 1981); theoretical aspects that determine conceptual coherence (Murphy & Medin, 1985) and conceptual change (Carey, 1982, 1985); and essential aspects that help account for people's beliefs about conceptual stability (Medin & Ortony, this volume; Smith,

Medin, & Rips, 1984). Although Smith and Medin (1981) once complained that conceptual cores did little to explain psychological data, it now seems that they are pulling their weight. The present suggestion is certainly consistent with these ideas, since cores, theories, or essences could easily be the source of many of the explanations subjects invoke in categorizing things. In fact, when categories have clear definitions, we would expect the explanations to amount to little more than a reference to the core. For example, it is natural to say that 794 is an even number *because* it's divisible by 2 (see Armstrong, Gleitman, & Gleitman, 1983).

However, the advantage of explanations can be seen most clearly when category decisions involve more than a recital of core properties. Recall, for example, Subject C's comments on why something 4.75 feet high would have to be a cereal box rather than a stop sign: "no one would see the stop sign if it were that small. . . . A stop sign would have to be a certain height, and while I wouldn't expect to see a cereal box that big, it wouldn't make sense to have such a small sign." This subject is not merely citing a core property of stop signs (as would be the case if he had said, "By definition, all stops signs are 7.5 feet high"). Instead, he reasons that a stop sign of that height would be hard to see and hence would not fulfill the function that stop signs are supposed to serve; therefore, the object probably isn't a stop sign. Information about a stop sign's function could certainly be among its core properties, but further inferences are necessary to use this information in classifying the object. The subject is obviously creating his argument on the fly to deal with the case at hand. The same can be said, I think, of Subject B's remarks on tennis balls and teapots, cited earlier. Of course, subjects do sometimes mention a core property and leave it at that, but this is not the only strategy available to them. Explanations have sufficient flexibility to subsume these strategies.

In short, resemblance theory seemed an attractive prospect: On this theory, similarity accounts for typicality, typicality measures degree of category membership, and degree of membership explains classification behavior. The problem is that this chain breaks somewhere in the middle, since neither similarity nor typicality fully accounts for degree of membership, as our subjects judge it. The view that classification is inference to the best explanation is not so tidy; it means that an adequate theory of categorization will have to await (or to develop alongside) an adequate theory of nondeductive reasoning. This may seem a disappointing state of affairs, but at least it locates categorization in the right space of complex mental processes.

NOTES

I have benefited from comments on an earlier version of this paper from audiences at the Workshop on Similarity and Analogy at the University of Illinois, and at colloquia at the Yale AI Lab, the University of Arizona Psychology Department, the UCSB Cognition Group, and the University of Wisconsin Psychology Department. Closer to home, Reid Hastie, Greg Murphy, Eldar Shafir, and Roger Tourangeau commented on an earlier draft of the manuscript. I should also like to acknowledge the help of Marshall Abrams, Judy Florian, and Janis Handte in conducting the experiments reported here. National Institute of Mental Health Grant MH39633 supported this research. The analyses of the second experiment were carried out at the University of Arizona, thanks to the Cognitive Science Committee and its chairman, Peter Culicover.

1 Murphy and Medin (1985) have a more general version of their relativity argument, which may not be susceptible to problems with the New Look. This is that similarity is always relative to some standard or set of criteria, where this standard varies from one category to the next. Thus, there is no way of computing similarity that is independent of the particular objects or categories in question, again leading to circularity when one tries to explain categorizing via similarity. (I have heard a very similar argument from Herbert Clark in discussion.) It is not clear to me, however, that this type of variation in the way similarity is computed entails that resemblance theory is viciously circular. Let us suppose, hypothetically, that people determine the similarity of an object to a category simply by counting the number of shared predicates in their representations. In general, if object O is compared to category C, a different set of shared predicates will be relevant than if O is compared to another category C′. And one might say, in this situation, that a different standard or set of criteria was involved for the O–C than for the O–C′ comparison. But although there is probably a lot that is wrong with such a theory, I do not think it would be fair to accuse it of circularity. The theory *would* be circular, of course, if it presupposed that people must know the category of which O is a member before they can carry out the similarity computation that is supposed to determine membership status. But it is hard to see why such a presupposition is necessary in our hypothetical case. In short, it seems you can have (some kinds of) relative standards without putting the resemblance theory into a loop.

2 An exception to this rule is Kahneman and Miller's (1986) norm theory. On this approach, categories are stored as remembered instances; however, the categorizing process can also make use of more abstract information that is computed from the instances (by a parallel process) at the time of the category decision. This clearly gives the model much more flexibility

than earlier exemplar theories and makes it less dependent on resemblance. Whether it escapes the problems with exemplar models that I describe later depends on how powerful the abstraction process is. It is certainly possible that subjects can compute on the fly properties like the variance or density of a category's distribution, which play a role in Experiments 1 and 2. But it is less likely that subjects obtain theoretical information (e.g., hidden biological properties, designers' intentions) in the same way. If these properties are crucial to category decisions (as in Experiments 3 and 4), then exemplars do not provide a rich enough representation.

3 In recent work, Lakoff (1987) proposes what he calls a prototype theory for a variety of linguistic and psychological phenomena. Although this theory is supposed to account for findings like those of Rosch and others, it is clearly much broader than the prototype models just described. In particular, similarity to a prototype has no special status in the theory. Instead, the basic representational unit is an *idealized cognitive model*, which can be thought of as an elaborated frame or schema. In terms of the typology of Figure 1.1, Lakoff's approach should be grouped with frame theories at the bottom of the scale, despite its title.

4 Douglas Medin suggested the psychophysical explanation to me in conversation. One problem with both the uphill/downhill and the psychophysics hypotheses is that one might well expect such factors to interact with the type of task subjects performed. Given the account just sketched for the main effect of task, variations in fixedness should produce a bigger difference for the Categorization group than for the Similarity group, whereas variations in subjective magnitude should yield a bigger difference for the Similarity group than for the Categorization group. Figure 1.2 shows a trend in the former direction, but the interaction is not a significant one. A larger experiment might be necessary to examine these possibilities in detail.

5 To get a measure of centrality for an interval, we computed the distance between the midpoint of the interval and median of the distribution. The distances for all the intervals within a given problem were then transformed to z scores in order to correct for differences in units of measurement across problems. (Recall that one problem involved temperatures Fahrenheit, another hair length in inches, and so on.) Finally, we multiplied each z score by -1 so that larger numbers represent intervals nearer the median. This gives centrality and frequency coefficients the same polarity in Table 1.1.

6 Much the same can be said of a second suggestion of Medin and Ortony (this volume): that both the Categorization and the Similarity subjects in the first study were computing similarity but with different instantiations of the mystery instance. This explanation is unlikely to be true, given subjects' reports of their own deliberations (as discussed later in this section). But even if it were, it would not be of much help to resemblance theory. We would still need an explanation of why instructions to categorize an instance caused subjects to imagine examples that differed systematically from the ones Similarity subjects envision. By hypothesis, the factors responsible for this difference cannot themselves be reduced to similarity and hence support the major claim of this chapter. This same point also applies to the experiments that I report in the next section.

7 Some subjects may, in fact, believe in properties that are necessarily true *de re*. An example might be Subject D's comments on decks of cards, cited earlier. For other subjects, however, the properties that are necessary for category membership are probably not ones they believe are essential to the objects' continued existence.

8 Susan Goldman suggested this way of putting the matter. Notice that I am not defending the view that *similarity* always means perceptual resemblance. We can obviously talk about the similarity of entities (e.g., ideas, goals, or personalities) that do not have perceptual attributes. Nor am I criticizing any particular theories of similarity. The experiments reported here, however, convince me that people place heavy emphasis on perceptual properties in their similarity judgments when this strategy is open to them. Perceptual resemblance may therefore be a root sense or default sense of similarity. This root sense could be extended in certain ways in order to describe the similarity of more abstract objects, but extending it in an unconstrained way – so that similarity can stand for any type of property comparison – may lead to the sorts of conceptual difficulties just discussed.

9 This way of looking at categorization is probably congenial to schema or frame theorists, since one of the points of these theories has been that a schema applies to an instance if it adequately accounts for the instance's properties. See Rumelhart and Norman (1988) for a review of these theories that stresses this perspective.

REFERENCES

Achinstein, P. (1983). *The nature of explanation*. Oxford: Oxford University Press.

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition, 13*, 263–308.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 629–654.

Brachman, R. J., & Schmolze, J. G. (1985). An overview of the KL-ONE knowledge representation system. *Cognitive Science, 9*, 171–216.

Brooks, L. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 169–211). Hillsdale, NJ: Erlbaum.

Carey, S. (1982). Semantic development: The state of the art. In E. Wanner & L. R. Gleitman (Eds.), *Language acquisition: The state of the art* (pp. 347–389). Cambridge: Cambridge University Press.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press (Bradford Books).

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*, 407–428.

Collins, A., & Rips, L. J. (in preparation). *An inductive approach to categorization*.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A

framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 234–257.
Goodman, N. (1970). Seven strictures on similarity. In L. Foster & J. W. Swanson (Eds.), *Experience and theory* (pp. 19–29). Amherst: University of Massachusetts Press.
Harman, G. (1966). Inference to the best explanation. *Philosophical Review, 74*, 88–95.
Harman, G. (1986). *Change in view: Principles of reasoning*. Cambridge, MA: MIT Press.
Hintzman, D. L., & Ludlam, G. (1980). Differential forgetting of prototypes and old instances: Simulation by an exemplar-based classification model. *Memory & Cognition, 8*, 378–382.
Jacoby, L. L., & Brooks, L. R. (1984). Nonanalytic cognition: Memory, perception, and concept learning. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 18, pp. 1–47). New York: Academic Press.
Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93*, 136–153.
Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–208). Cambridge: Cambridge University Press.
Keil, F. C. (1986). The acquisition of natural kind and artifact terms. In W. Demopoulos & A. Marras (Eds.), *Language learning and concept acquisition* (pp. 133–153). Norwood, NJ: Ablex.
Lakoff, G. (1987). *Women, fire, and dangerous things: What categories tell us about the nature of thought*. Chicago: University of Chicago Press.
McCloskey, M., & Glucksberg, S. (1979). Decision processes in verifying category membership statements: Implications for models of semantic memory. *Cognitive Psychology, 11*, 1–37.
Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207–238.
Miller, G. A., & Johnson-Laird, P. N. (1976). *Language and perception*. Cambridge, MA: Harvard University Press.
Murphy, G. L., & Medin D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289–316.
Oden, G. C., & Lopes, L. L. (1982). On the internal structure of fuzzy subjective categories. In R. R. Yager (Ed.), *Recent developments in fuzzy set and possibility theory* (pp. 75–89). Elmsford, NY: Pergamon.
Osherson, D. N., & Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition, 9*, 35–58.
Quine, W. V. O. (1969). Natural kinds. In W. V. O. Quine, *Ontological relativity and other essays* (pp. 114–138). New York: Columbia University Press.
Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Erlbaum.
Rosch, E. & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573–605.
Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance, 2*, 491–502.
Rumelhart, D. E., & Norman, D. A. (1988). Representation in memory. In

R. C. Atkinson, R. J. Herrnstein, G. Lindzey, & R. D. Luce (Eds.), *Handbook of experimental psychology* (Vol. 2, pp. 511–587). New York: Wiley.
Schank, R. C., Collins, G. C., & Hunter, L. E. (1986). Transcending inductive category formation in learning. *Behavioral and Brain Sciences, 9*, 639–686.
Schwartz, S. P. (1980). Natural kinds and nominal kinds. *Mind, 89*, 182–195.
Smith, E. E. & Medin, D. L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.
Smith, E. E., Medin, D. L., & Rips, L. J. (1984). A psychological approach to concepts: Comments on Rey's "Concepts and stereotypes." *Cognition, 17*, 265–274.
Smith, E. E., Osherson, D. N., Rips, L. J., & Keane, M. (in press). Combining concepts: A selective modification model. *Cognitive Science*.
Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review, 81*, 214–241.
Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327–352.
Tversky, A., & Gati, I. (1978). Studies of similarity. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 79–98). Hillsdale, NJ: Erlbaum.
White, T. H. (1954). *The bestiary: A book of beasts, being a translation from a Latin bestiary of the 12th Century*. New York: Putnam.