Psychological Review

VOLUME 92 NUMBER 3 JULY 1985

The Role of Theories in Conceptual Coherence

Gregory L. Murphy
Brown University

Douglas L. Medin
University of Illinois

The question of what makes a concept coherent (what makes its members form a comprehensible class) has received a variety of answers. In this article we review accounts based on similarity, feature correlations, and various theories of categorization. We find that each theory provides an inadequate account of conceptual coherence (or no account at all) because none provides enough constraints on possible concepts. We propose that concepts are coherent to the extent that they fit people's background knowledge or naive theories about the world. These theories help to relate the concepts in a domain and to structure the attributes that are internal to a concept. Evidence of the influence of theories on various conceptual tasks is presented, and the possible importance of theories in cognitive development is discussed.

Why is a given set of objects grouped together to form a category? That is, why is it that some groupings are informative, useful, and efficient, whereas others are vague, absurd, or useless? The current surge of interest in people's concepts has provided much information about conceptual structure and content. Yet, the central question of what makes a category seem coherent has only been sketchily addressed and incompletely answered.

A somewhat unusual, but nonetheless useful, example arises from an old puzzle of biblical scholarship, the dietary rules associated with the abominations of Leviticus, which produce the categories *clean animals* and *unclean animals*. Why should camels, ostriches, crocodiles, mice, sharks, and eels be declared unclean, whereas gazelles, frogs, most fish, grasshoppers, and some locusts be clean? What could chameleons, moles, and crocodiles have in common that they should be listed together? That is, what is there about clean and unclean animals that makes these categories sensible or coherent?

The main thesis of this article is that current ideas, maxims, and theories concerning the structure of concepts are insufficient to provide an account of conceptual coherence. All such accounts rely directly or indirectly on the notion of similarity, and we argue that the notion of similarity relationships is not sufficiently constraining to determine which concepts will be coherent or meaningful. These approaches are inadequate, in part, because they fail to represent intra- and inter-concept relations and more general world knowledge. We propose a different approach in which attention is focused on people's theories about the world.

The keystone of our explanation is that people's theories of the world embody conceptual knowledge and that their conceptual

This research was supported by United States Public Health Service Grant MH32370 (to DLM) and by National Science Foundation Grant 83-15145 (to GLM). This article is a fully collaborative venture, and the order of authorship is arbitrary.

The authors wish to acknowledge the helpful comments of Lawrence Barsalou, Maureen Callanan, Eve Clark, Sarah Hampson, Reid Hastie, Robert Macauley, Barbara Malt, Glenn Nakamura, Andrew Ortony, Elissa Newport, Brian Ross, Ed Shoben, Richard Shweder, and Ed Smith on an earlier draft.

Requests for reprints should be sent to Douglas Medin, Department of Psychology, University of Illinois, 603 East Daniel, Champaign, Illinois 61820, or to Gregory Murphy, Hunter Laboratory of Psychology, Brown University, Providence, Rhode Island 02912.

organization is partly represented in their theories. At one level, this statement is trivially true: For example, one's understanding of chemistry influences one's concept of substances like *water.* It would be very odd for a person to believe, for example, that water is animate, and yet to understand the phase relations between water, ice, and steam. Surely there is some consistency between people's concepts and their understanding of interacting objects and forces in the world, but the connection between the two has very seldom been spelled out. We attempt to specify the connection between theoretical and conceptual knowledge and to recast conceptual theory in that light.

Current theories of conceptual structure, including those we have proposed ourselves, represent concepts in ways that fail to bring out this relation between conceptual and theoretical knowledge. For example, one theory treats concepts[1] as exemplars organized around a central prototype (see B. Cohen & Murphy, 1984; Osherson & Smith, 1981). It is difficult to see how these concepts might be related to or constrained by one's knowledge of the world. Another influential model (actually, a set of models) treats concepts as collections of features of some sort (see Smith & Medin, 1981).[2] Although this model may be broad enough to involve theoretical knowledge, it does not particularly promote it, nor does it suggest what concepts people are likely to have and why. In particular, the features suggested by most theories of concepts have excluded the theoretical connections we will discuss.

In this article, we do not propose a new model of conceptual representation. Rather, we present a theory of what the glue is that holds a concept together and an account of what sorts of concepts are easy to learn, use, and remember, with the understanding that conceptual models must build appropriate structures to account for the facts discussed.

When we argue that concepts are organized by theories, we use *theory* to mean any of a host of mental "explanations," rather than a complete, organized, scientific account. For example, causal knowledge certainly embodies a theory of certain phenomena; scripts may contain an implicit theory of the entailment relations between mundane events; knowledge of rules embodies a theory of the relations between rule constituents; and book-learned, scientific knowledge certainly contains theories. Although it may seem to be glorifying some of these cases to call them theories, the term connotes a complex set of relations between concepts, usually with a causal basis. Furthermore, these examples are similar to theories used in scientific explanation (Achinstein, 1968). Later on, we offer a list of some general properties of people's theories and review examples illustrating the utility of thinking of concepts as being embedded in theories.

The philosopher W. V. O. Quine was one of the first to make a case for the use of theories in determining category membership. In his classic article, "Natural Kinds," Quine (1977) argued for both a psychological and a societal progression from an innate, similarity-based conception of kinds to a theoretically oriented, more objective basis. Whereas early societies could only depend on perceptual and functional qualities to differentiate objects into classes, modern society can use techniques of chemical, physical, and genetic analysis in order to classify. Quine further argued that, in a true case of ontogeny recapitulating phylogeny, modern children begin with innate, perceptually based similarity metrics to define their kinds, only to have them successively replaced by scientific knowledge (to the limits of their education and our scientific progress). As Quine (1977, p. 171) puts it:

> One's sense of similarity or one's system of kinds develops and changes and even turns multiple as one matures, making perhaps for increasingly dependable prediction. And at length standards of similarity set in which are geared to theoretical science. This development is a development away from the immediate, subjective, animal sense of similarity to the remoter objectivity of a similarity

[1] Many authors do not clearly distinguish between *concepts* and *categories*. We use *concepts* to refer to mental representations of a certain kind, and *categories* to refer to classes of objects in the world. Past writers seem to have used category to mean the mental representation of a class of objects, or both the representation and the objects themselves. However, this distinction is important to account for deviations between the two, as when someone's concept of *animal* does not actually include all animals.

[2] Throughout this article, we use the terms *feature, attribute,* and *property* interchangeably.

determined by scientific hypotheses and posits and constructs. Things are similar in the later or theoretical sense to the degree that they are interchangeable parts of the cosmic machine revealed by science.

Although we do not subscribe to Quine's claims about societal progression (or the view that the use of scientific theories is necessarily more objective), we agree with his conclusion that one's theories explicate the world and differentiate it into kinds. We also concur with him that the notion of similarity must be extended to include theoretical knowledge. Although we focus on explicit theories as a source of conceptual coherence, it is likely that a broader view of theoretical knowledge will be needed to provide a complete account. People use some kinds of theoretical knowledge implicitly, only becoming aware of doing so when confronted with a mismatch or failure of that knowledge (as may arise in cross-cultural contact). Furthermore, even people's explicit theories may often not reach the rigor and consistency expected from a scientific theory (Nisbett & Ross, 1980; A. Tversky & Kahneman, 1980). Thus, the kind of theory Quine had in mind (an explicit, scientific one) is too narrow to fully explain coherence. The next section reviews previous approaches to conceptual coherence and their limitations.

Approaches to Conceptual Coherence—The Insufficiency of Similarity

We have already hinted at what we mean by a coherent category. It is one whose members seem to hang together, a grouping of objects that makes sense to the perceiver. We do not give an operational definition of coherence because we do not wish to tie it to a particular theoretical framework. There are a number of measures that might reflect coherence, including how easily the concept is learned and used, and there may be others that are not known yet.

It is important to distinguish this notion of coherence from the related one of *naturalness,* as used by Keil (1981) and others. Natural concepts are said to be those formed out of basic ontological categories, such as *living thing* or *intelligent being.* For example, a category that included only thoughts and fish would cross ontological boundaries improperly and would therefore form an unnatural concept. However, as we show later, a concept that is unnatural (according to this definition) may be coherent because people have some theory that it plays a part in. In short, most of people's concepts are probably natural and coherent, but the issue of what makes a concept hang together cannot be solved solely by recourse to such ontological categories.

Perhaps the most powerful explanation of conceptual coherence is that objects, events, or entities form a concept because they are similar to one another. The basic idea is that objects fall into natural clusters of similar kinds (that are dissimilar to other clusters), and our concepts map onto these clusters. Thus, similarity may be the glue that makes a category learnable and useful. Although it is true that category members seem similar, Quine (1977) pointed out that using similarity as the basis for concepts may raise the very questions it was meant to answer. Without some explanation of why things seem similar, we are left with an equivalent problem; many things appear to be similar just because they are members of the same category. In more practical terms, estimates of similarity may be influenced by people's knowledge that the things being compared are in the same (or different) categories.

To use a rough analogy, winning basketball teams have in common scoring more points than their opponents, but one must turn to more basic principles to explain why they score more points. In the same way, similarity may be a by-product of conceptual coherence rather than its determinant—having a theory that relates objects may make them seem similar. Goodman (1972, p. 437) goes so far as to say, "Similarity, ever ready to solve philosophical problems and overcome obstacles, is a pretender, an imposter, a quack. It has, indeed, its place and its uses, but is more often found where it does not belong, professing powers it does not possess."

We shall argue that, at its best, similarity only provides a language for talking about conceptual coherence. Certainly, objects in a category appear similar to one another. But does this similarity explain why the category was formed (instead of some other) or its ease of use? Suppose we follow A. Tversky's (1977) influential theory of similarity, which defines it as a function of common and distinctive features weighted for salience or importance. If similarity is the sole explanation of category structure, then an immediate problem is that the similarity relations among a set of entities depend heavily on the particular weights given to individual features. A barber pole and a zebra would be more similar than a horse and a zebra if the feature "striped" had sufficient weight. Of course, if these feature weights were fixed, then these similarity relations would be constrained. But as Tversky (1977) demonstrated convincingly, the relative weighting of a feature (as well as the relative importance of common and distinctive features) varies with the stimulus context and experimental task, so that there is no unique answer to the question of how similar one object is to another. To further complicate matters, Ortony, Vondruska, Jones, and Foss (1984) argued persuasively that the weight of a feature is not independent of the entity in which it inheres. The situation begins to look very much as if there are more free parameters than degrees of freedom, making similarity too flexible to explain conceptual coherence.

A further major complication derives from the fact that no constraints have been provided on what is to count as a feature or property in analyses of similarity. Suppose that one is to list the attributes that *plums* and *lawnmowers* have in common in order to judge their similarity. It is easy to see that the list could be infinite: Both weigh less than 10,000 kg (and less than 10,001 kg, . . .), both did not exist 10,000,000 years ago (and 10,000,001 years ago,. . .), both cannot hear well, both can be dropped, both take up space, and so on. Likewise, the list of differences could be infinite. Furthermore, there are some attributes that are true of only a small number of the category members—perhaps there are some orange plums or some lawnmowers run by robots. What is the cutoff for excluding attributes that are not universal, or must they all be included (Murphy, 1982a)? The point is that any two entities can be arbitrarily similar or dissimilar by changing the criterion of what counts as a relevant attribute. Unless one can specify such criteria, then the claim that categorization is based on attribute matching is almost entirely vacuous (see Goodman, 1972).

These arguments about attributes fly in the face of perceptual experience that seems to naturally partition at least some entities into categories. Of course, there are some categorizations that blatantly contradict perceptual similarity (e.g., categorizing whales as *mammals*), which indicates that one's theories can override or at least select from perceptual information. Yet, it is true that the perceptual system has some built-in constraints on what will count as an attribute and which attribute relations are salient (see Ullman, 1979, for elegant work that gets at some of these constraints). The problem with the abstract notion of similarity is that it ignores both the perceptual and theory-related constraints on concepts, when in fact they are doing most of the explanatory work. How much of our conceptual system is based on perceptually determined features and how much on theoretical features has yet to be determined. In general, people seem to be flexible about similarity (even perceptual similarity), and we know relatively little about nonperceptual constraints. Thus, we attempt to provide part of the answer to how people choose relevant attributes for concepts and how they weight those attributes in their conceptual processes. However, we wish to reduce the importance of individual attributes in conceptual representations and to emphasize the interaction of concepts in theory-like mental structures.

We now consider some candidate principles for category coherence that rely directly or indirectly on the notion of similarity. We begin by considering some standard maxims about what makes a good category and then turn our attention to particular categorization theories and their implications for category structure. Finally, we examine the widespread assumption that category judgments are based on some form of attribute matching that maps directly onto similarity. There are serious problems and limitations associated with each of these principles.

The Insufficiency of Similarity-Based Measures of Category Structure

Although we have already argued that similarity does not sufficiently constrain concepts, it may be that there are some general processing principles that are based on similarity that have greater explanatory power. For example, there is considerable evidence that the most useful concepts are neither the most specific nor the most abstract, but are at an intermediate level of abstraction (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Although we would not want to equate concept coherence with these basic level concepts, such concepts are obviously highly coherent. Finding a metric that picks out these intermediate level categories is nontrivial.

Rosch and her colleagues argued that basic-level categories maximize cue validity (Rosch, 1978; Rosch et al., 1976), the conditional probability that an object is in a category, given that it has some cue (or attribute) associated with the category. A coherent category should have many such cues, whereas a poor category has only inconsistent cues, or very few good ones. Categories with the highest cue validity would be expected to be particularly useful in perceptual categorization. Unfortunately, this measure incorrectly predicts that superordinate (i.e., the most inclusive) categories are always more coherent than any of their subordinates, inasmuch as anything that cues membership in one category also cues membership in its superordinates. For example, if something has feathers, it is likely to be a bird, but it is at least equally likely to be an animal. (See Murphy, 1982a, 1982b, for details and consideration of similar measures.)

Perhaps coherent or useful categories are the ones that allow the most inferences to be made—after all, one purpose of categories is to enable inferences that may not be apparent from individual exemplars. If an object is a *dog,* for example, one can infer that it has ears, barks, has fur, and so forth, even if those properties have not been observed, whereas a vague category like *thing* or *object* enables few if any inferences to be made. Actually, this measure, which could be called *category validity,* is the reverse of cue validity, as it might be represented as the conditional probability that something has various attributes given its category membership. Accordingly, it has the reverse problem: Medin (1983) noted that the more specific a category, the more inferences it allows—individual objects being the limiting case for which one can specify the greatest number of correct "inferences."

It may well be possible to find measures that pick out intermediate levels of abstraction. For example, some weighting function combining cue validity and maximizing inferences surely would (e.g., Jones, 1983). But even here, there is little ground for confidence that we can measure coherence formally because the basic level appears to change with expertise (e.g., Rosch et al., 1976). One could reflect such changes by adding features or modifying feature weights, but again, these additions and modifications are doing the explanatory work. Similarity may be able to describe such facts, but it does not explain them.

The Insufficiency of Correlated Attributes

Another organizing principle for categories is the notion of correlated attributes. Rosch et al. (1976; Rosch, 1978) proposed that natural categories divide the world up according to clusters of features, that they "cut the world at its joints." That is, attributes of the world are not randomly spread across objects, but rather appear in clusters. Furthermore, basic categories (which are the most useful and efficient) are said to maximize the correlational structure of the environment by preserving these attribute clusters.

Another motivation for the correlated attributes principle is the idea that organisms are constantly "going beyond the information given" to draw inferences and make predictions. For example, on the basis of seeing a round object in a gymnasium, one might predict with considerable confidence that it would bounce (though this inference would be wrong in the case of a medicine ball). In general, these predictions or inferences prove to be accurate to the extent that people correctly perceive such attribute correlations.

This *correlational structure* account implies that some version of the similarity models considered above is correct at a descriptive level because categories develop to group objects with a cluster of features and to exclude objects with different features. Yet, this account also makes a stronger claim than do those previous models: It is not undifferentiated similarity that holds a concept together, but some more elaborated structure of correlations. In this sense, the correlated attributes principle is deeper than are general notions of similarity. That is, an organism programmed to take advantage of attribute correlations will tend to form categories that have high within-category and low between-category similarity as a *consequence* of detecting correlations.

One problem with the correlated attributes notion is that there are so many possible correlations that it is not clear how the correct ones get picked out (see Keil, 1981, for an elaboration of this point). It would seem that some additional principle is needed to provide further constraints on category cohesion (e.g., perhaps correlations are more readily noticed if the parts are spatially contiguous or subserve the same function). A cause and its effect may be highly correlated, but they would probably be placed in different categories. Another problem is that the mental representation of correlated features needs to be specified further, including a specific mechanism that results in their making concepts more coherent.

We will not criticize this account because we believe that concepts that preserve correlations are in fact more coherent. However, we also believe that there are further principles that explain this fact—that correlated attributes do not provide a full account of conceptual cohesiveness. To anticipate our later arguments, we believe that feature correlations are partly supplied by people's theories and that the causal mechanisms contained in theories are the means by which correlational structure is represented.

The Insufficiency of Categorization Theories

Smith and Medin (1981) divided theories of category representation into three basic approaches: the classical view, the probabilistic view, and the exemplar view. It is natural to ask whether these theories imply useful constraints on concept or category goodness. For the most part, they do not.

Classical view. The classical view has it that categories are defined by singly necessary and jointly sufficient features. The major problems with this view as a structural principle are that many categories may not conform to the classical view (see Medin & Smith, 1984; Mervis & Rosch, 1981; Smith & Medin, 1981, for reviews) and, equally seriously, that defining attributes do not ensure coherence. This theory does not pick out some defining feature sets as better or more appropriate than others. For example, a category consisting of striped things that have more than one leg and that weigh between 11 and 240 kg satisfies a classical view definition, but does not seem sensible or cohesive.[3]

Probabilistic view. The probabilistic view denies that there is a common core of criterial properties and argues that concepts may be represented in terms of features that are typical or characteristic, rather than defining. First, we should note that the criticism just made for the classical theory applies here as well: Without supplementation, the probabilistic view cannot tell which combinations of features form possible concepts and which form incoherent ones. It would not rule out the following combination of typical features: bright red, swims, has wings, eats mealworms, is found in Lapland, and is used for cleaning furniture. Clearly the mere fact that this combination is probabilistic does not mean that it is coherent (see Murphy & Wisniewski, 1985).

Second, many processing models associated with the probabilistic view have the general constraint that the summary representation coupled with appropriate processing assumptions should accept all members and reject all nonmembers. The formal term for the constraint that categories be partitionable on the basis of a summing of evidence (i.e., the presence of features) is that the categories be separable by a linear discriminant function (Sebetsyen, 1962). That is, categories should

[3] One might argue that this concept does not seem coherent simply because few objects actually contain all these features. (This objection could also apply to our first criticism of the probabilistic view below.) However, other empty concepts are fully coherent; in fact, our culture is full of fictional or mythical concepts that are perfectly coherent without having any members. The classical view does not explain why some empty categories seem reasonable and others do not. Furthermore, if we could provide a context in which our example *became* coherent (e.g., perhaps a stage prop with those characteristics is needed), the classical view would have nothing to say about this change.

be separable on the basis of a weighted, additive combination of their features: Categories that are not linearly separable should be difficult to learn and use.

Is linear separability important for actual concepts? One way of evaluating its importance is to set up two categorization tasks that are similar in major respects, except that in one task the categories are linearly separable and in the other categorization task they are not. Although this question has not received much attention, what little evidence there is is negative. In a series of four experiments varying instructions, category size, and stimulus materials, Medin and Schwanenflugel (1981) found no evidence that linearly separable categories were easier to learn than categories that were not linearly separable. Thus, linear separability does not appear to be a necessary property of "good" concepts.

Exemplar view. The exemplar view agrees with the probabilistic view in holding that concepts need not have criterial properties and, further, claims that categories may be represented by their individual exemplars rather than by some unitary description of the class as a whole (see Medin & Schaffer, 1978). Obviously, such a view offers no principled account of conceptual structure because it does not constrain what exemplars are concept members. Although most exemplar theories assume that category members are similar, we have already argued that this alone is not a full explanation of coherence.

In brief, it seems that none of the three major views of category representation provides a principled account of category cohesiveness.

General Insufficiency of Attribute Matching and Similarity

Our claim is not only that approaches to category coherence based on similarity have to date been unsuccessful, but that, in principle, they will prove to be insufficient. We see three major problems with an exclusive focus on similarity and the associated practice of breaking concepts into constituent attributes or components: First, it leads naturally to the assumption that categorization is based solely on attribute matching; second, it ignores the problem of how one decides what is to count as an attribute; and third, and more generally, it engenders a tendency to view concepts as being little more than the sum of their constituent components. All of these problems derive directly or indirectly from failing to view concepts in terms of the relations between exemplar properties and the categorization system: Human interests, needs, goals, and theories are ignored.

Categorization as attribute matching. Our objection to the idea that categorization derives from attribute matching is that it may prove to be too limited. For example, the attributes associated with higher level concepts may be more abstract than those of lower level concepts or exemplars. Instead of attribute matching, categorization may be based on an inference process (see Collins, 1978). For example, jumping into a swimming pool with one's clothes on is, in all probability, not associated with the concept *intoxicated,* yet that information might well be used to decide that a person is drunk. That is, categorizing the person as intoxicated may explain his or her behavior, even though the specific behavior was not previously a component of the concept. This inference process must be fairly complex, taking into account the context: In our example, the behavior could imply drunkenness in one context and heroism in another (e.g., jumping into the pool to save someone from drowning). Concepts may represent a form of shorthand for a more elaborate theory, and a concept may be invoked when it has a sufficient explanatory relation to an object, rather than when it matches an object's attributes.

A major respect in which attribute matching may be too limited is that our representations may include information concerning operations, transformations, and (indirectly) relations among attributes (see also Hampton, 1981). Much of our reasoning about concepts may be based on contraints about operations that are permissible. Consider the following situation:[4] Suppose that all the soda cans you have come into contact with have been 7.5 cm in diameter and that all the silver dollars

[4] The example is based on an idea provided by Lance Rips.

you have seen have been 4.0 cm in diameter. Suppose further that you are told that some entity has a diameter of 5.0 cm and you are asked whether it is more likely to be a soda can or a silver dollar. To our minds, it is more likely to be the can. One reason for this guess is that we know that silver dollars are mandated by law to be a particular size, whereas soda cans just happen to be of a uniform size. Alternatively, one might have made the opposite conjecture based on the knowledge that soda cans have to be a particular size to fit soda machines, whereas there is little reason for the particular size of silver dollars (other than in casinos). The point is that, whichever choice is made, it clearly does not derive solely from attribute matching or size similarity judgments, but rather from our knowledge about transformations and operations associated with concepts, and this, in turn, relies heavily on our general world knowledge.

This case could be recast as an example of attribute matching in which the attributes are higher order properties. For example, one's concept of *silver dollar* could have the attribute "used in machines sensitive to exact size." Although this is technically true, it misses the important point that the explanatory work is again being done by the theory-constrained processes that generate these complex attributes, rather than by attribute matching per se. Thus, although attribute matching could be made to be consistent with these facts, it does not explain or predict them by itself.

Although we believe that theoretical factors are important in people's categorizations, it seems likely that people can develop automatic routines for identifying objects as members of concepts when the concepts have consistent perceptual features. For example, one probably does not usually invoke much theoretical knowledge in categorizing something as a *robin*. The main influence of theories on perceptual categorization may be on novel objects and borderline cases, and when the categorization must be justified or explained. In short, we emphasize the theoretical aspects of categorization, but we do not mean to exclude the use of primarily perceptual information. Current research on categorization gives evidence that both are important (Kelter et al., 1984; Murphy & Smith, 1982).

Selecting attributes. Frequently, attributes are treated as givens or at least as sufficiently transparent that all one has to do is to ask experimental subjects to list them. As we have noted, this largely ignores the problem of what can count as an attribute. The formal models of category coherence mentioned above gain credence from their precise formulation of coherence, but they have no precise way in which to choose or exclude the attributes that form their basis.

More recently, some work has begun to be directed at this issue. Barsalou and Bower (1983), for example, showed that two types of properties are likely to be activated during processing. First, properties that have high diagnosticity may be active inasmuch as they are useful for distinguishing instances of a conept from instances of other conepts. Second, properties relevant to how people typically interact with instances of a concept are likely to be frequently active (see also Barsalou, 1982, for further arguments). Note that forms of typical interaction themselves vary with context (see Roth & Shoben, 1983).

Barsalou and Bower's (1983) research reinforces our thesis that the explanatory work is on the level of determining which attributes will be selected, with similarity being at least as much a consequence as a cause of conceptual coherence. In addition, their reference to typical interactions with objects suggests the causal schemata and scripts that we have said are important in conceptual representations. The properties that distinguish concepts may be greatly determined by people's goals, which are linked to their theories about the objects.

Concepts as equivalent to their components. The more general problem associated with viewing concepts as equivalent to the sum of their components has a long history. Consider the following quote from John Stuart Mill (1843/1965):

> The laws of the phenomena of the mind are sometimes analogous to mechanical, but sometimes also to chemical laws. When many impressions or ideas are operating in the mind together, there sometimes takes place a process of a similar kind to chemical combination. When impressions have been so often experienced in conjunction, that each of them calls up readily and instantaneously the ideas of the whole group, those ideas sometimes melt and coalesce into one another, and appear not several ideas but one; in the same manner as when the seven prismatic colors are presented to the eye in rapid succes-

> sion, the sensation produced is that of white. But in this last case it is correct to say that the seven colors when they rapidly follow one another *generate* white, but not that they actually *are* white; so it appears to me that the Complex Idea, formed by the blending together of several simpler ones, should, when it really appears simple, (that is when the separate elements are not consciously distinguishable in it) be said to *result from*, or be *generated by*, the simple ideas, not to *consist of* them. . . . These are cases of mental chemistry: in which it is possible to say that the simple ideas generate, rather than that they compose, the complex ones. (p. 29)

Although many investigators would agree that mental chemistry is a more apt metaphor for understanding concepts than is mental composition, the core of this distinction does not appear to have taken hold. Again, one would have thought that mental chemistry would convey a concern with *relations* (and constraints associated with them), *operations*, and *transformations* on components, as opposed to an exclusive focus on components (i.e., features) as independent entities.

One defense of the attribute-matching perspective is that relations and operations themselves might be treated as attributes. To take this step, however, is to concede that attributes may have a complex internal structure. Relations need arguments, and arguments and relations mutually constrain one another. This internal structure means that one is working with more than a list of simple attributes and that constraints and explanatory power will derive from this richer structure.

It also seems likely that the listing of category attributes, although helpful for certain methodological uses (e.g., Rosch & Mervis, 1975), may drastically underestimate people's categorical knowledge, because part of their knowledge is about relations of category features to each other and of category members to the world. Thus, a person who simply *memorized* the attributes of some categories without knowing more about the object domain might have very different concepts than does a person with elaborated theories. These differences would show up in the uses of categories in language understanding, naming, problem solving, and other situations (some described below), but perhaps not in feature listings.

Summary of the Two Approaches

In our discussion, we have lumped together a number of accounts of concept representation and categorization under the general heading of *similarity-based approaches to concepts*. Although they differ in many respects, these accounts have in common the characteristic that they treat concepts as collections of attributes. In our critique of this approach, we argued that it is *insufficient* to explain conceptual coherence and the richness of conceptual structure. (In later sections we review more empirical data on this issue.) We emphasize *insufficient* here because we do not want to imply that this approach is completely wrong or misleading. It is clear that category members seem similar to one another, but we have argued that similarity is too flexible to give any specific, natural explanation of conceptual coherence. One could see our approach as supplying the constraints missing from the similarity explanation, rather than simply contradicting it.

Table 1 summarizes the differences of the similarity-based approach and the theory-based approach on a number of dimensions (some of which we have yet to address). The entries for the similarity-based approach uses *attribute* as a general term for features, propositions, and other simple chunks of knowledge. Under the theory-based approach, *underlying principle* is used to refer to the causal connections, script links, and explanatory relations that we have been invoking as parts of theories.

In general, it can be seen that the similarity-based approach requires a minimum of conceptual organization and relations, whereas the theory-based approach emphasizes both. One way to describe this difference is to say that the theory-based approach expands the boundaries of conceptual representation: In order to characterize knowledge about and use of a concept, we must include all of the relations involving that concept and the other concepts that depend on it. To explain conceptual coherence, the processes that operate oñ a concept must be considered in addition to the information directly stored with it.

Concepts as Embedded in Theories

We have no illusions about having solved the problem of concept coherence. Unless one can specify constraints on what a theory is, it may not help at all to claim that conceptual coherence derives from having a

Table 1
Comparison of Two Approaches to Concepts

Aspect of conceptual theory	Similarity-based approach	Theory-based approach
Concept representation	Similarity structure, attribute lists, correlated attributes.	Correlated attributes plus underlying principles that determine which correlations are noticed.
Category definition	Various similarity metrics, summation of attributes.	An explanatory principle common to category members.
Units of analysis	Attributes.	Attributes plus explicitly represented relations of attributes and concepts.
Categorization basis	Attribute matching.	Matching plus inferential processes supplied by underlying principles.
Weighting of attributes	Cue validity, salience.	Determined in part by importance in the underlying principles.
Interconceptual structure	Hierarchy based on shared attributes.	Network formed by causal and explanatory links, as well as sharing of properties picked out as relevant.
Conceptual development	Feature accretion.	Changing organization and explanations of concepts as a result of world knowledge.

theory. Table 2 lists five general properties that many theories manifest, along with some suggested roles that these properties may play in thinking about conceptual coherence. Because theories are flexible, conceptual coherence may also be. For example, the category *apple-or-prime number* does not appear to be a very coherent concept. In our view, this lack would derive mainly from the lack of clear internal or external structure in a theory about such a category. The relations that apples participate in (e.g., eating, biological relations) overlap very little with the relations that prime numbers participate in.

One could develop a scenario, however, in which this category might make sense.[5] For example, suppose that one of our colleagues in the math department, Wilma, has only two interests: prime numbers and apple farming. We might, then, form the concept *prime numbers-or-apples,* which is explained as "topics of conversation with Wilma." This explanation provides very little structure, however, so that it would probably be less coherent than the concept *apples-or-oranges.* By adding more explanatory links, one could make the concept more coherent. For example, one could try to explain why Wilma has only those two interests. Through reference to naive personality theory and by exploring the properties of apples and prime numbers, one could elaborate a theory about why a person would have just these interests. If this theory were consistent with one's other world knowledge, then it would also supply external structure to the concept. Whether this concept could ever become very coherent is an open question, depending on the status of the theory itself and the plausibility of competing theories. The point is that one might have a theory that could connect (to some degree) objects that seem to share very few features.

The rest of this article can be viewed as an amplification of the entries in Table 2 and in the right half of Table 1. In the following sections, we discuss how considering theories improves on the simple similarity accounts of these issues.

The Role of Theories in Cognition

Our claim is that representations of concepts are best thought of as theoretical knowledge or, at least, as embedded in knowledge that embodies a theory about the world. In this section, we reconsider some of the issues raised in the previous section and show how the addition of theoretical knowledge fills many of the gaps in explaining conceptual coherence.

[5] Larry Barsalou helped to develop this example.

Table 2
General Properties of Theories and Their Potential Role in Understanding Conceptual Coherence

Property of theories	Speculation about role in conceptual coherence
"Explanations" of a sort, specified over some domain of observation.	Constrains which properties will be included in a concept representation. Focuses on certain relationships over others in detecting feature correlations.
Simplify reality.	Concepts may be idealizations that impose more structure than is "objectively" present.
Have an external structure—fit in with (or do not contradict) what is already known.	Stresses intercategory structure. Attributes are considered essential to the degree that they play a part in related theories (external structures).
Have an internal structure—defined in part by relations connecting properties.	Emphasizes mutual constraints among features. May suggest how concept attibutes are learned.
Interact with data and observations in some way.	Calls attention to inference processes in categorization and suggests that more than attribute matching is involved.

Theories and Attribute Selection

Earlier we raised the issue of what is to count as an attribute. One answer is to rely on consensual validation: If several experimental subjects list some property as an attribute of some concept, then that attribute is included in the concept. Rosch and Mervis (1975) have shown that these listed attributes can be used to predict goodness of example ratings and times to verify that an exemplar is a member of a category (see Mervis & Rosch, 1981, for a review).

Although this technique has generated important data for theories of categorization to explain, we may wish to consider the question of how people choose attributes to list. One might think that participants can simply retrieve the most important features of the target concept and report them. However, there are reasons to believe that the process of generating attributes is more complex.

First of all, most of the research involving attribute listing employs judge-amended tallies. The reason for this is that participants may list attributes at one level of abstraction and fail to include them at a lower level of abstraction. For example, they may list "two-legged" for *bird*, but not for *robin, eagle,* and other specific birds. B. Tversky and Hemenway (1984) analyzed this behavior in terms of cooperative rules of communication (Grice, 1975) and implicit contrast sets (e.g., "two-legged" does not distinguish between *robin* and *eagle,* and so it may not be listed). The idea of implicit contrast sets may also explain why "does not fly" is much more likely to be listed for *penguin* than for *rainbow trout.* Thus, the subject's conception of the relevant contrast set, as well as the desired level of specificity, influences the choice of which features to list. It appears, then, that attribute listings may be quite constrained by factors that are only beginning to be studied.

We submit that attribute listings and the representations behind them are further constrained by the theories that the categories are involved in. Subjects list not everything they know about a concept, but rather those features that are particularly salient and diagnostic in their background knowledge (and that seem most relevant in the situation, as B. Tversky & Hemenway, 1984, noted). For example, most people realize, upon reflection, that the attribute, "flammable," applies to wood, money, certain plastics, and (sadly) even animals. Yet, it probably would be found only in the conceptual representation (and the listings) for the first of these categories, presumably because of the known role of wood in human activities. Some attributes are prominent in our concepts because of their importance in our other knowledge about the world, and others are excluded because of their irrelevance to our theories. The concept *money* is central to our theories of economic and social interaction, in which the attribute of flammability plays no role. Thus, it is apparently not part of our representation of *money* even though it may easily be inferred as true of most money.

Miller and Johnson-Laird (1976) also noted the importance of theories in specifying attributes of lexical concepts. They contrast a concept's *core,* which contains theory-based attributes, with attributes that are perceptually salient and therefore useful in identification, but with little connection to the intrinsic nature of the concept. They describe the concept's core as being "an organized representation of general knowledge and beliefs about whatever objects or events the words [in a lexical field] denote—about what they are and do, what can be done with them, how they are related, what they relate to" (p. 291). They make the explicit equation: "A conceptual core is an inchoate theory about something" (p. 291). Although it is often difficult to draw the line between core features and more peripheral features, Miller and Johnson-Laird's description emphasizes the importance of external and internal structure of a concept's features in the core.

Theories and Correlated Attributes

We raised the possibility earlier that coherent concepts have clusters of correlated features. We then raised the question of how conceptual representations take advantage of these clusters. In other words, what is the difference between representations of categories with feature correlations and those without feature correlations that result in the former being more coherent than the latter?

Smith and Medin (1981, pp. 84–86) discussed two possibilities. One is to represent correlated features as one single feature. For example, the features "flies," "has wings," and "has a beak" might be combined into one global feature. Smith and Medin pointed out that this solution is unprincipled and counterintuitive, in that the compound feature really corresponds to three independent features that must be separated in other representations (e.g., bats and penguins have only two of the three features). The other possibility they mentioned is to link and label features that are correlated. So, all three pairs of the above features would have arcs labeled CORRELATED connecting them.[6] This has more intuitive appeal—its main drawback being the explosion of feature links it would engender—and Smith and Medin tentatively accept it.

This feature-linking solution has computational tractability. It can adequately represent feature correlations that might be accessed by processes using the concept. However, this solution misses an important insight. Features in categories are not correlated by virtue of random combinations. Rather, correlations arise from logical and biological necessity: Animals and artifacts have structural properties in order to fulfill various functions, so that some structural properties tend to occur with others, and certain structures occur with certain functions. It is no accident that animals with wings often fly or that objects with walls tend to have roofs. Even less obvious correlations, such as the one between furniture being made of wood and also having a flat top (Malt & Smith, 1984), usually have clear explanations.

Suppose that people are not only sensitive to feature correlations, but that they can deduce *reasons* for those correlations, based on their knowledge of the way the world works. Perhaps, then, the connection between those features is not a simple link, but a whole causal explanation for how the two are related. For example, one can connect "has wings" to "flies" by one's intuitive knowledge of the use of wings to support a body on air pressure; "has walls" and "has a roof" are connected by their common function of protection from the elements. This approach avoids the explosion of CORRELATED links because it draws on previously existing knowledge about the attributes to connect them: The links are already in memory. Furthermore, memory research has shown

[6] The links would not have to be labeled as CORRELATED—they might simply be associations that simultaneously activate two features, and this pattern of activation could be used to infer that the features are correlated. That is, the correlations might be computed rather than specifically stored. However, this version is also subject to the objections we raise to the more explicit representation of correlated attributes.

that it is difficult to remember correlated facts through simple associations; when the facts are tied together by a theme of previous knowledge, memory interference is reduced (Bower & Masling, 1978; Day & Bellezza, 1983; Smith, Adams, & Schorr, 1978).

Medin, Altom, Edelson, and Freko (1982) found in experiments with novel categories that people are, in fact, sensitive to feature correlations and that they use them in their categorization judgments (see also L. B. Cohen & Younger, 1983; Younger & L. B. Cohen, 1984). This was true even when overall typicality was controlled for. Thus, people do spontaneously use feature correlations to aid their judgments. Notably, during the debriefing, participants frequently offered reasons for *why* the correlation was present. They were not simply computing correlations but were developing and using theories to explain the correlations and to structure the concept.

Theories and Concept Use

So far, we have argued on theoretical grounds that people's concepts must be integrally tied to their theories about the world. A large part of this discussion has been somewhat abstract, dealing with various measures of conceptual coherence and accounts of category structure. This approach to conceptual coherence also has empirical implications for concept *use.* Although many process models of concept use involve attribute matching or similarity judgments, we argue that a number of lines of research give evidence of the use of causal knowledge, rules, theoretical consistency, and other theory-like knowledge. This section reviews evidence pertaining to how theories are involved in specific uses of concepts.

Correlated Attributes

We have already suggested that theories are necessary for people to explain feature correlations. Medin et al. (1982) showed that people are sensitive to empirical correlations of features in their category judgments, as Rosch et al. (1976) suggested they should be. However, features that are correlated in people's mental representations may not always reflect empirical relations in the world, but may derive instead from people's theories about the relations between the features. Although these theory-driven relationships may actually exist, people may never have empirical data to confirm or disconfirm their expectancies. Examples of these feature pairs are amount of education and income, zodiac sign and personality, rate of speech and intelligence, and amount of rehearsal and strength in long-term memory. Again, we rush to point out that some of these pairs may be truly correlated, but others probably are not. The property that they have in common is that they are predicted by (some) people's theories about the world, rather than being suggested by observation. In fact, some of them are so theory laden that it would be difficult to see how one could detect them without the theory to direct measurement. When a correlation is perceived to exist on the basis of one's theories, but has no basis in empirical fact, it is called an *illusory correlation.*

Chapman and Chapman (1967, 1969) presented evidence that therapists and naive subjects using certain psychodiagnostic tests perceived correlations between test results and psychological disorders when in fact there were none—or even when the opposite correlation obtained. They concluded that people's expectancies prevented them from objectively evaluating the relation between the test and mental illness. Other studies have confirmed the effects of theories on perception of correlations, although not always to the same degree (Crocker, 1981; Wright & Murphy, 1984). Bower and Masling's (1978) research suggested that the important factor may be that people be able to construct a causal explanation for a correlation, rather than that it match their current knowledge. Murphy and Wisniewski (1985) provided some preliminary evidence that theory-based correlations are actually used to form conceptual representations.

One could imagine a case opposite to the illusory correlation one, in which the observer perceived a correlation but could find no explanation for it; there might be no way to connect the two attributes in one's mental scheme of things. One of us (DLM) has recently completed a set of studies in which people were asked to sort descriptions of entities into categories. For example, in one case, the descriptions were symptoms and the categories were hypothetical diseases. The task was set up so that people could sort on the basis of two different sets of correlated attributes. The two sets of correlated attributes differed in terms of how readily people might think of a causal association between them. Although people are flexible enough that they can link many pairs of symptoms, pilot work suggested that it is easier to link some pairs (e.g., dizziness to earaches, and weight gain to high blood pressure) than others (e.g., dizziness to weight gain, or earaches to high blood pressure). People showed a strong tendency to cluster on the basis of correlated attributes for which a causal link could readily be made. Furthermore, subjects mentioned such linkages to justify their sorting. For example, they might say that an ear infection could disturb the vestibulary organ and produce both dizziness and earaches. Thus, feature correlations may be important in conceptual representations primarily when they can be represented as theoretical knowledge.

There is also evidence that a prior theory can facilitate perception or learning of contingencies and correlations. For example, in processing numerical information involving possible correlations, performance may be improved dramatically simply by the addition of meaningful labels for the variables that suggest their theoretical significance (e.g., Adelman, 1981; Camerer, 1981; Miller, 1971; Muchinsky & Dudycha, 1974; Wright & Murphy, 1984). Camerer (1981) showed that people could learn an interaction between variables when they were labeled in accordance with prior beliefs (i.e., factors thought to affect wheat futures in the commodity market), but failed to learn when the same problem was given as an abstract task involving arbitrary labels.

Linear Separability in Categorization

We mentioned earlier that linear separability does not appear to be a natural constraint on human categorization. One reason for this may be that people's theories, and hence their categories, typically have more internal structure than can be captured by an independent summing of evidence or by similarity to a prototype. If this is true, then if a prior theory suggests that summing or similarity matching is appropriate, linear separability may in fact become important for categorization.

Recent work by Wattenmaker, T. Murphy, Dewey, Edelson, and Medin (1984) supported this idea. In one study the descriptions were properties of objects, and the categories were structured such that the typical attributes for one category would all be desirable properties if one were searching for a substitute for a hammer (e.g., flat surface, easy to grasp). In one condition subjects were given the notion of hammer substitutes, and in another condition they were not. The idea was that a hammer would act as an ideal standard and that subjects could judge how similar examples were to the hammer prototype (through independent summing of features).

When prior theories were developed or suggested, linearly separable categories were in fact easier to learn than were nonlinearly separable categories. The reverse held when no theory was suggested. This result depends on the theory evoked being compatible with a summing of evidence. By suggesting a different form of theory, one should be able to reverse this pattern of results. For example, if one category corresponded to psychologists, one might discourage people from summing up component information by alerting them to the fact that there are both experimental and clinical psychologists and that their traits may differ considerably. The attribute "likes computers" might predict category membership for experimental but not clinical psychologists. In a close analogue of this example, Wattenmaker et al. (1984) found a differential facilitation in learning categories that were not linearly separable.

The point of these examples is quite simple. One cannot describe some abstract conceptual structure as simple or complex, independent of the form of theory that might be brought to bear on it. When theory and structure match, the task becomes simple; when there is a mismatch between theory and structure, the task becomes difficult.

Theories and Prototype Structure

Assuming that most concepts have a typicality structure, people must discover this structure when they learn a new concept.

When they encounter a new object, they must judge how typical it is of a variety of concepts. Both of these tasks may require use of a theory. Barsalou's (1983, in press) research on goal-derived categories presents a particularly clear example in which theories are crucial to deriving conceptual structure. He investigated categories such as *things to do at a convention.* He found, first, that people are less likely to discover that four objects are in one of these categories when they do not know the goal that relates them (Barsalou, 1983, Experiment 4). Second, he showed that the typicality structure of goal-derived categories was not simple family resemblance (similarity of the category members), but rather how well each instance satisfies the goal (Barsalou, in press). The reader may wish to introspect on what the category is that includes the objects children, jewelry, portable TVs, paintings, manuscripts, and photograph albums. Furthermore, which of the items mentioned is the most typical? Because the objects have low family resemblance, the task is nearly impossible. However, once the theme *taking things out of one's home during a fire* is known, these judgments become easy. Notice that this concept is not a "natural" one according to the criteria given by Keil (1981), yet it does seem to hang together in its context. Such examples suggest that theories can elucidate the relations among very different objects and thereby form them into a coherent category, even if they do not form a "natural" class.

A third interesting aspect of Barsalou's (in press) research involves some comparisons he made between natural and goal-derived concepts. In the process of showing that the exemplars of goal-derived categories had typicality ratings that correlated with the degree to which they satisfied the relevant goal, Barsalou performed similar computations on common concepts. Although the underlying dimensions for natural categories were speculative (e.g., for *fruit,* how much people like it), they proved to be significantly correlated with exemplar goodness even after the effects of frequency and family resemblance had been partialed out. This observation suggests that natural concepts may be partly organized in terms of underlying dimensions that reflect how the concept normally interacts with people's goals and activities.

Fillmore (1982) made a related suggestion about the source of typicality structures. He argued that lexical concepts are represented in terms of *idealized cognitive models.* For example, the concept *bachelor* can be defined as an unmarried adult male, in the context of human society in which certain (idealized) expectations about marriage and marriageable age are realized. The existence of "poor examples" of this concept—for example, Catholic priests, homosexual men, men cohabiting with a girlfriend—does not mean, Fillmore argued, that the concept itself is ill-defined. Rather, the claim is that the idealized cognitive model does not fit the actual world perfectly well. An entity may deviate from the concept (i.e., may be atypical) either because it fails to satisfy "unmarried, adult male" or because the idealized model is imperfectly realized. Clearly, such a model is an example of what we have been calling *theories,* inasmuch as it provides a means of connecting many concepts in order to explain diverse facts. Mohr (1977) argued that this is the correct way to view Platonic universals, and Lakoff (1982) developed this notion of idealized models in some detail.

In this view, the relation between concepts and exemplars is analogous to the relation between theory and data. Not only may data be somewhat noisy, but theories also typically involve simplifying assumptions that trade parsimony for power. As Kuhn (1962) argued, theories depend on a particular background of accepted beliefs and assumptions that is taken for granted—until contradictory data begin to accumulate. Fillmore's (1982) point was that categorizing objects also depends on background assumptions about the world, and our concepts have developed in the context of those assumptions. To some degree, then, it may be these simplified models that give rise to unclear cases, and when anomalous or unclear cases arise, our background assumptions become more salient.

We may underestimate the importance of implicit theories or background assumptions about the world because of their very implicitness. Ziff (1972) provided some delightful examples of the importance of implicit conceptual schemes in understanding. For example, it seems sensible to say "a cheetah can outrun a man." But what about a 1-day old cheetah, or an aged cheetah with arthritis, or a healthy cheetah with a 100-pound weight on its back? What we mean when we say that a cheetah can outrun a man is that under some tantalizingly difficult-to-specify conditions, a cheetah would outrun a man. Ziff referred to this set of conditions as a conceptual scheme and made the point that two people understand each other to the extent to which these conceptual schemes are shared. These implicit theories heavily constrain our understanding of relations among concepts.

Expertise

The prevailing view of expertise with regard to concepts seems to be that experts differ from novices primarily in making finer distinctions (as implicitly expressed by Dougherty, 1978; Rosch et al., 1976). In that view, experts have many more specific categories than do novices, and they see those categories as being very distinct. It has often been suggested that experts should have different concepts from novices, but few studies have actually investigated their conceptual structure. Much of the relevant work has involved cross-cultural comparisons in anthropological studies of lexical structure (e.g., Berlin, Breedlove, & Raven, 1973; Dougherty, 1978; others are cited by Mervis & Rosch, 1981). For example, members of agricultural societies are experts on plants and animals and have many names for specific animal concepts, whereas Berkeley undergraduates are novices and have few such names (Dougherty, 1978; Rosch et al., 1976).

However, there may well be differences between experts and novices besides the *amount* they know about a category and the *number* of categories they can differentiate. Certainly, experts have better developed theories about the domain than do novices. How would this affect their conceptual structure? A reasonable null hypothesis would be that experts simply know more: They have more information about each category, and they know more categories. Although these quantitative predictions seem likely, we do not believe that they are the only differences. Experts in some domain probably know more relations between the objects in the domain. They can see connections where novices notice none because their theories lead them to look for certain similarities, regularities, and cause-effect relations. For example, biologists notice crucial similarities between shrimps, moths, grasshoppers, spiders, and crabs, putting them together in one class (the arthropods). We assume that naive observers would make more pragmatic distinctions, probably separating the flying, crawling, and water-living animals. The biologist's theories of evolution and physiological structures express themselves in the concept of the arthropods and would come into play explicitly when categorizing unfamiliar objects.

There is increasing evidence for the view that experts make far-reaching connections that affect their concepts, in addition to having greater specific knowledge. Murphy and Wright (1984) examined the concepts of experts and novices in child psychopathology. The novices were college undergraduates with no experience in abnormal psychology. Three other groups ranged in expertise from beginning counselors at a summer camp for disturbed children to clinical psychologists with extensive experience in the field. All of the subjects listed attributes of the three major categories of emotionally disturbed children. Surprisingly, experts' concepts were not more distinctive—in fact, the more expert the subjects, the more their categories seemed to overlap.

This result is somewhat counterintuitive because experts in clinical psychology are expected to classify people into different groups, and the more distinctive their concepts of the groups, the easier this would be. This finding points out that classification is not the only purpose for concepts. Like all psychologists, these experts wanted to find *explanations* for behavior, and those explanations point out commonalities to all cases of child psychopathology (analogous to the zoologist's search for organizing features in biological classifications). For example, the professional psychologists listed "feels angry" and "feels sad" for all categories, presumably because of their theories about the motivational and cognitive concomitants of psychopathology. Novices also have theories of psychopathology, but they are apparently more superficial, accounting for surface differences between the categories. For example, they listed "feels sad" as an attribute of depressed children only, and "feels angry" exclusively for aggressive children.

One interpretation of these findings is to attribute them to the fuzziness or even invalidity of psychopathological categories. However, similar evidence was reported in the realm of physics problems by Chi, Feltovich, and Glaser (1981), who noticed that novices classify physics problems using "surface features" that are only roughly correlated with physical principles. Experts, on the other hand, apparently categorized problems on the basis of the major principles used in their solutions. Consequently, "experts are able to 'see' the underlying similarities in a great number of different problems, whereas novices 'see' a variety of problems that they consider different" because the surface features differ (Chi et al., 1981, p. 130). As a result, the experts made fewer, larger classes than did the novices. Chi et al.'s results also highlight the fact that similarity is in the eyes—and theories—of the beholder.

It seems safe to assume that the physicists' classifications were not simply fuzzier than the novices' (as one might argue for the clinical psychology case). Similarly, the biologist's class of *arthropods* is accepted as valid, even though it is much more inclusive than preferred novice concepts (see Berlin et al., 1973; Rosch et al., 1976). These examples provide evidence that people's theories may lead them to form concepts that they would not normally have and to alter the content of other categories.

Cross-Cultural Research

An intriguing possible implication of the approach we have proposed has to do with cross-cultural differences in concepts. Clearly, people in different cultures have different theories about the world, which should cause them to have different concepts. In fact, there are a number of tantalizing examples of cultural differences in classification tasks (see the review by Cole & Scribner, 1974). One well-documented culturally dependent phenomenon is the assignment of the basic level of categorization. Rosch et al. (1976) first noted that the basic level of their American subjects was more general than that of people from agricultural, nonindustrial societies (as described by Berlin et al., 1972). Dougherty (1978) and Geoghegan (1976) discussed these differences in depth and suggested that domains that are important to a culture are more fully individuated and elaborated both in the language and conceptual system. The basic level is more specific in such domains than in others. Such cultural dependence is evidence against the idea that the basic level is purely determined by features in the environment. In our view, this happens because the greater salience of a domain promotes more elaborate knowledge structures in the domain, which in turn can differentiate more specific concepts.

However, these differences in salience do not exhaust the effects of cultural knowledge on concepts. One example is that the Karam of New Guinea do not consider a cassowary a bird. Bulmer (1967) argued that this is not merely because the cassowary does not fly, but because of its special role as a forest creature and its resulting participation in an elaborate antithesis in Karam thought between forest and cultivation. This antithesis is further related to basic concerns with kinship roles and rights. Apparently, the Karam's theories about forest life and cultivation produce different classifications than do our culture's biological theories. (For other similar examples, see Luria, 1976; Tambiah, 1969; and the review by Cole & Scribner, 1974.) For categories that are more conceptual than perceptual, cross-cultural differences may be even more evident. Shweder and Miller (in press) demonstrated the importance of cultural presuppositions in social categories involved in person perception, in a strong parallel to the position of this article.

Linguistic Innovations and Complex Concepts

Because people's representations of word meanings are probably closely tied to their concepts (see E. Clark, 1983), our theory should also have implications for semantic interpretation. This influence can probably best be seen in the understanding of innovative uses of language, which require modification of existing word meanings in order to be interpreted. A similar problem is the formation of complex concepts, in that existing concepts must be modified in order to create a new meaning.

Clark and Clark (1979) discussed the creation and interpretation of denominal verbs, which are often innovative—created for a single use by a particular speaker—rather than conventional like most word uses. Examples include *Max teapotted the dean,* and *the boy porched the newspaper,* in which the concepts *teapot* and *porch* must be modified to produce verb interpretations. To explain how people understand such innovations, Clark and Clark referred to people's "generic theories" of objects: their physical characteristics, ontogeny, and potential roles. For example, one's knowledge of boys, newspapers, and porches allows one to conclude that *the boy porched the newspaper* refers to throwing a paper on the porch (rather than making it into a porch or pasting it on the porch). The same denominal verb in a different sentence frame would involve a different interpretation, as in *the builder porched the house.* People's conceptual knowledge is heavily involved in producing and constructing interpretations of such sentences, and that knowledge apparently includes the origins and usual roles of such objects, as we have argued.

Combining simple concepts into compound concepts may involve similar processes.[7] For example, how does one generate the concept *pet fish* from the concepts *pet* and *fish?* One possibility is the "classical" method of set intersection (Osherson & Smith, 1981). For example, *pet fish* would be formed by taking the intersection of all things that are *pets* and all things that are *fish.* Much of the early concept acquisition literature assumes such an account.

Unfortunately, this view has a great deal of trouble with many complex concepts. Consider, for example, *ocean drive, expert repair,* or *horse race.* These concepts are not intersective at all. *Ocean drives* are not both *oceans* and *drives; horse races* are not both *horses* and *races.* Linguists discussing nominal compounds have argued that the meaning of these terms is determined by a *mediating relation* between the two nouns (Kay & Zimmer, 1976), but there is no single relation that will construct any complex concept (see Adams, 1973). For example, a *horse race* is a race of horses, but an *ocean drive* is not a drive of oceans. An *expert repair* is a repair done by an expert, but an *engine repair* is probably not a repair done by an engine. So, no single relation (like set intersection) can describe all or even most compound concepts. Furthermore, the construction of complex concepts is not a simple operation on the features of the two concepts, such as feature overlap or projection. Although some of the features of *finger* get carried over onto *finger cup,* considerable knowledge is needed to specify which features are affected and how they are combined with the features of *cup.* Whenever people form complex concepts or understand compound nouns, they must be using their background knowledge of the way the world works in order to create the correct concept. In short, the formation of complex concepts requires mental chemistry rather than the simple addition of components.

B. Cohen and Murphy (1984) argued that it is impossible to explain how people form such compound concepts using only *knowledge independent* operations. That is, they said that it is impossible to say in advance what a complex concept *XY* means knowing only the meaning of *X* and *Y*, but that extensive knowledge relating *X* and *Y* comes into play in order to arrive at just the right compound. In the context of our discussion, this point translates into the use of people's implicit theories and operations on concepts. For example, one's knowledge of the use of vehicles, their parts and what they do, and mishaps that happen to them can lead one to combine *engine* and *repair* to get "repair of an engine." One's knowledge about *experts* leads one to combine *expert* and *repair* differently. The interpretation of a compound concept may be thought of as a hypothesis generated by background theories.

[7] It is difficult to give operational criteria to separate simple from complex concepts. One clue is whether the concept has a single-word name or requires multiple words (Berlin et al., 1973). Yet, some compound noun phrases name unitary concepts, for example, *washing machine.* Rather than argue for an operational distinction here, we have used simple and complex concepts that are intuitively clear: The simple concepts are described by a single word, and they combine to form apparently complex concepts.

Related Ideas

The notion that people's concepts are tied up with their theories is not totally new to psychology (note the earlier discussion of Miller & Johnson-Laird, 1976). Rumelhart (1980) made a related analogy in describing his theory of knowledge representation. Schemata, he suggested, are like theories in that they embody expectations of what things co-occur and how properties are related (pp. 37–38). Unfortunately, the actual schemata he presented are not rich enough to express people's knowledge about those relations and co-occurrences. For example, the schema for *buy* includes agents, an object being sold, the transfer of money, and so forth, which expresses a simple theory about financial transactions. However, people's full understanding of buying events includes information about the motives and desires of the seller and buyer, expectations about the relation between the money and the purchase (that they should be of near-equivalent worth), and a number of legal and cultural requirements. Our intent here is not to criticize Rumelhart's representations: It is possible that a complete schematic representation could contain all the necessary theoretical knowledge, especially when the relations among various schemata are included. Our point is that the full knowledge people have about concepts goes beyond that normally given in such discussions.

In memory research, the shift from emphasis on memory traces (the Ebbinghaus tradition) to processes of memory construction and reconstruction (the Bartlett tradition) has been well documented. Whereas early memory researchers investigated the passive laying down and decay of traces, more recent investigators have posited active encoding and reconstructive processes (Bransford, Barclay, & Franks, 1972; Cofer, 1973; Jenkins, 1974). These processes are based on the relation of the material to the rest of the knowledge base, rather than on abstract learning rules.

In the area of judgment and inferences, A. Tversky and Kahneman (1980) considered the specific place of causal knowledge in decision making, implicating it in a number of judgment situations. Other work suggested that people give great weight to their theories about people and the world relative to statistical evidence (see Nisbett & Ross, 1981; Wright & Murphy, 1984, for reviews). In particular, abstract rules of judgment and decision making (e.g., Bayes's theorem or Luce's choice axiom) apparently do not characterize people's decisions. Although this field has engendered much controversy (e.g., L. J. Cohen, 1981), it seems clear that people use specific theories of the world, sometimes inappropriately, to make predictions and decisions.

In the area of language comprehension, people's use of theoretical knowledge has been reflected in two ways. First, there has been increasing interest in people's theories of communication itself (although this factor is not usually described in this way). Grice (1975) first pointed out that speakers and hearers use their beliefs about the purposes of a conversation in order to make and understand implications. H. Clark and Carlson (1982) and H. Clark and Murphy (1982) discussed how listeners and readers use their beliefs about the purposes and methods of communication to understand reference and various aspects of meaning. In essence, these discussions have dealt with how implicit theories of communication come into play in everyday language use (we have already mentioned that they may affect the listing of concept attributes). Second, psycholinguists have begun to emphasize how people's knowledge of the discourse topic allows them to fully understand the discourse. In this case, people use their theories about the domain being discussed to rule out anomalous interpretations and to resolve ambiguities and vagaries. Simple models of lexical decomposition and inference no longer seem adequate to the task of explaining the range and depth of language understanding—see Collins, Brown, and Larkin (1980), Johnson-Laird (1981), Rumelhart (1981), and Schank and Abelson (1977).

Finally, the area of problem-solving has embraced the notion of mental models in people's reasoning about complex systems (see articles in Gentner & Stevens, 1983). Content-free reasoning strategies such as means-ends analysis or logical deduction seem unable to account for the relative difficulty of different problems or for individual differences. Instead, investigators have suggested that subjects form a simplified mental model of a system and simulate its behavior in order to make a prediction or evaluation. Clearly, the subject's theory about the system and the domain it operates in will greatly determine his or her problem solution. Furthermore, concepts in the domain are determined to a great extent by the whole model in which they operate.

Although the psychological domains we have discussed are disparate, there is a clear theme running through them. In each case, a simple model based on invariant principles of organization or process has been found too inflexible to account for human abilities. People appear to use content-specific knowledge or theories to process information and to represent new knowledge. The importance of these constructive, knowledge-based processes appears to be well established for these fields.

It is interesting to note that procedural approaches to categorization from artificial intelligence have sometimes depended on theory-like structures. For example, the sorting algorithm that seems to best capture people's free sorting of entities into categories is not an exclusively bottom-up processor (Michalski, 1983; Michalski, Stepp, & Diday, 1981). Rather, the basic procedure of Michalski's program operates on the level of *descriptions* of clusters and aims to maximize criteria having to do with what represents a good description. These criteria include such factors as simplicity, the fit between descriptions and the entities, and a bias for conjunctive descriptions. Therefore, a good description can be thought of as having the character of a good theory (the former is a consequence of the latter).

Philosophy of science has long considered the question of whether concepts are integrally bound up with theories. Unfortunately, there is little agreement on the answer, with opinions ranging across the extremes. Philosophers such as Kuhn and Feyerabend argued that scientists with different theories about a domain must have different concepts in the domain, even if their concepts have the same names. For example, physicists who held the wave theory of light had concepts of *light, color,* and the like that differed from those of physicists who held the particle theory. Other philosophers have downplayed this possibility or have argued that any such conceptual differences are usually insignificant; Suppe (1977) provides a complete discussion of both sides. Although this issue remains controversial, it does seem clear that present-day scientific concepts are quite different from past understanding of the same concepts as a result of new theories and knowledge. Current work in philosophy of science focuses on the boundaries of such conceptual differences.

Conceptual Development

The study of children's concepts and semantic development may be a crucial area for showing the importance of theories in conceptual structure. Not only do children lack words for many of the entities, events, and situations that adults have words for, they may have quite different theories about how those entities, events, and situations are related. Although there is still no consensus on children's cognitive and linguistic representations, we believe that some of the accepted findings speak to the issues we have raised.

The most influential theory of semantic development in recent years has been Eve Clark's Semantic Feature Hypothesis (E. Clark, 1973a, 1973b; Richards, 1979). Following accepted linguistic analyses, Clark used sets of components or features as semantic representations. She suggested that children's first semantic representations of a word are a subset of the adult features (although occasionally completely incorrect features will sneak in) and that development consists primarily of adding features as they are learned. The Semantic Feature Hypothesis successfully described much of the data, including the order of acquisition of words in many domains and common naming errors (see E. Clark, 1973a, 1973b, 1983).

For a variety of reasons, this theory is no longer widely accepted in its original form (see Carey, 1982; E. Clark, 1983; Richards, 1979), a trend that is consistent with our previous arguments about the insufficiency of feature-based models of concepts. It is not our purpose to review the literature in semantic development here, but we would like to highlight the studies that shed light on how theories might influence conceptual development and that contrast with the featural view.

One of the first studies was E. Clark's (1973b) demonstration of nonlinguistic "biases." Previous data had suggested that children learned locative prepositions in the order, *in, on, under.* For some time, they treated *under* as if it meant *in* or *on.* One explanation for these data was that all three words had the same semantic representation at first, and that with increasing experience, children added features to differentiate them. However, Clark showed that children had biases about spatial arrangements that influenced their performance in comprehension tasks. That is, if told, "Put the block under the crib," they might put it in the crib instead, because of their knowledge of usual spatial relations. In fact, they made the same error even when imitating nonverbal actions. Clark suggested that the youngest children tested (about 21 months old) know only that *in, on* and *under* are spatial terms and that they use spatial strategies to respond to those words. Children depend on their knowledge of supporting surfaces and containers, and the usual orientation of objects to interpret utterances with locative prepositions (see also H. Clark, 1973). In a sense, they are depending on implicit theories of spatial relations to understand and learn new words. Semantic development, therefore, consists of coordinating one's conceptual knowledge with the conventions of word use. As E. Clark (1973b) remarked, in this view it becomes very difficult to determine when a child knows the correct meaning of a word: One must try to access linguistic knowledge separately from the conceptual basis, which may be impossible in practice.

Carey (1982) also provided a critique of the notion of feature accretion as an explanation of semantic development. The acquisition of spatial adjectives like *big, little, tall, short, thick,* and *thin* had been taken to be evidence for the Semantic Feature Hypothesis: *Big–little* were analyzed as having relatively few semantic components, *tall–short* as having additional features specifying orientation, and *thick–thin* as having yet more features (see below). The order of acquisition followed this analysis. Carey, however, argued that the difficulty of learning *thick–thin* was not the mere number of features it contains, but rather that it requires attending to "theory-laden" features specifying that the dimension being referred to is "tertiary." In order to resolve the meaning of these terms, Carey claimed that children must learn the complex spatial system we use in our culture to assign such spatial adjectives, and that this system is not part of their beginning theories about the world. Presumably, the learning of this spatial system goes hand in hand with learning the language. We would add that the child must also have extensive background knowledge about individual objects in order to determine their primary and tertiary dimensions. This knowledge is necessary to interpret the use of *thick* when applied to objects as diverse as doors, lines drawn on a page, people, and bicycle tires.

Keil and Carroll (1980) provided a demonstration that children do not represent spatial terms as abstract features, but that their understanding of them was inextricably bound up in their knowledge of the world. They demonstrated that children's willingness to describe something as *tall* depended on what they believed they were naming. A child might be able to pick out the tallest of a trio of mountains, but not the tallest of a trio of blanket piles—even though the same picture was used for both. Keil and Carroll proposed that the children had not yet extracted the abstract meaning of *tall,* but they did know some things that *tall* is used to describe (e.g., people, houses, mountains). Until they learn the full meaning, they depend on some primitive theory of what tall things are like.

The work of Ellen Markman and her colleagues (see Markman & Callanan, 1984) is also suggestive in this context. It is known that young children have difficulty learning and using superordinate concepts (Horton & Markman, 1980; Mervis & Crisafi, 1982; Rosch et al., 1976), which is not surprising, given their loose structure. Presumably, it is difficult for children to infer the functional relationship that often characterizes superordinates (*furniture, tool, vehicle, weapon,* etc.). Callanan & Markman (1982) suggested that 2- and 3-year-old children understand superordinates not as *classes,* but as *collections* of objects. That is, rather than thinking of *furniture* as a name that applies to individual objects, they think of it as a name for a group or configuration of a number of objects. They may believe that *furniture* refers to an arrangement of chairs and couches around a table in the living room. However, children do not seem to have the same problem with most basic concepts, which are much more perceptually based (Callanan & Markman, 1982).

These results are consistent with the interpretation that children cannot simply memorize that couches, chairs, tables, and bureaus are all *furniture*—they seem to need an explanation for this grouping, which might otherwise be incoherent. For them, the most reasonable explanation may be a spatial configuration rather than the more abstract functional explanation that adults use. If this is true, then it demonstrates the importance of underlying relationships in learning concepts. (See Gentner, 1983, for a similar claim concerning analogical transfer.)

Finally, in considering children's errors in learning noun and verb meanings, Carey (1982) argued that children's problems arise not from faulty linguistic abilities, but rather from an impoverished conceptual structure. For example, to fully understand a word like *buy* may require a sophisticated understanding of monetary exchange. But children may interpret *buy* merely as "get at a store." More generally,

> The components revealed by semantic analyses of the adult lexicon cannot be expected to be the primitives over which the child forms his hypotheses about the meanings of words. Often those components are theoretical terms in theories the child has not yet encountered, and they therefore require theory building on his part before they are available to his conceptual system. (Carey, 1982, p. 374)

Of course, the relation cuts both ways: An impoverished conceptual structure might prevent someone from learning a word fully, but in other cases, language learning influences the conceptual structure. A child may learn about monetary exchanges through learning the meaning of *buy* and *sell* rather than through direct experience or lessons in economics. As the child learns about the distinction between *buy, sell, trade, give,* and so forth, he or she learns complex concepts that are central to understanding society.

In her own studies of biological concepts (as described in Carey, 1982), Carey followed the development of concepts like *animal* and *living thing.* She attempted to empirically test Quine's theory that an innate similarity metric is replaced by a scientific metric as the basis of concepts. She did find some evidence for such a shift; children first organize properties of animals around their applicability to humans, but later develop a more systematic organization based on biological functions. However, even the youngest children (4 years old) showed some use of biological knowledge in their categorizations. Adults and children both rated a toy monkey as being more similar to people than a worm was. However, adults and children also agreed that the worm was more likely to have a spleen than was the toy monkey (a spleen was described as "a green thing inside people"). Apparently, even the youngest children differentiated surface similarity from category membership. Although worms may be less similar to people than are toy monkeys, they are more similar *in some respects,* namely, common biological functions. Carey's results demonstrate that it is those respects that determine category membership, rather than similarity as a whole. As Carey (1982, p. 386) put it, "The child's rudimentary biological knowledge influences the structure of his concept *animal* in several ways, even for children as young as 4. To that extent, *animal* functions as a natural kind concept by Quine's characterization."

A crucial question that arises in considering theories in conceptual development is when they make their first appearance. One might argue that children form their first concepts through perceptual similarity; then, as they learn more about the world, they incorporate knowledge into their concepts, where it has increasing importance. On this view, the similarity-based accounts of coherence are correct for early concepts, at least, to the extent that we can ascertain built-in constraints on the perception of similarity. The question, then, is just when theories begin to have an effect. Our view is that theories are important very early: E. Clark's (1973b) results showed that children under 2 years old demonstrated a variety of spatial biases. Other researchers have found that very young children can distinguish the sorts of objects that receive proper names from those that do not, presumably reflecting a theory of individuality (Gelman & Taylor, 1984; Katz, Baker, & Macnamara, 1974). As we argued earlier, these biases and preconceptions may be biologically determined to some extent through perceptual and cognitive structures (see H. Clark, 1973; Keil, 1981). Although young children may not have scientific theories or sophisticated schemata, they may well use their understanding of their world, or prototheories, in forming concepts (see Karmiloff-Smith & Inhelder, 1974/1975, for more direct evidence). Rather than a shift from similarity-based concepts to more theoretically-based concepts, perhaps all concepts are integrated with theories, but children's theories change radically.

Some studies of infants' categories have shown prototype structures in children a few months old (e.g., L. B. Cohen & Younger, 1983). The age of the children and the structure of the stimuli leave little doubt that the infants are forming concepts based on perceptual similarity. However, as we have already noted, similarity itself is not an unanalyzable relation, and perceived similarity also changes with development (see Kemler, 1982). It is certainly possible that children's prototheories of the functions, relations, and importance of objects have effects quite early. Exactly when they do is an empirical question, one that we hope will get some attention.

The Classical Theory of Concepts

A major bone of contention in the theory of concepts has been the question of whether concepts can be specified by necessary and sufficient features. Wittgenstein (1953) sparked the debate among philosophers, which continues today among psychologists and linguists as well. Although this classical theory appeared to be dead (see, e.g., Smith & Medin, 1981), a number of hybrid theories have arisen. Osherson and Smith (1981), for example, suggested that the conceptual core is all or none, and that prototypes and other nonessential information about a concept are used mainly for identification, but are not strictly part of the concept. McNamara and Sternberg (1983) argued for a mixed theory, in which concepts are represented by both defining (necessary and sufficient) and characteristic features.

We do not mean to resolve the philosophical issues here. Regardless of one's theory of concepts, it is a fact that most people believe that there are necessary and sufficient features that define concepts. McNamara and Sternberg (1983) documented this fact convincingly, and informal questioning reveals that naive subjects are loathe to admit that there are no truly defining features, even when they cannot produce any (Rosch & Mervis, 1975). Armstrong, Gleitman, and Gleitman (1983) asked subjects whether they thought certain categories were all or none or had graded membership. For their natural categories, the percentage of subjects who responded "all or none" ranged from 24% (for *vehicle*) to 71% (for *sport*). People apparently have a strongly held belief that there are defining attributes for categories, in spite of the failure of psychologists, linguists, and philosophers to find any. (Suggestions for necessary features have been made, but these never seem to define the concept sufficiently; e.g., perhaps all *trips* involve motion, but this does not separate them from innumerable other events.) What we will try to explain is, where do these beliefs come from?

A natural prediction from our previous discussion is that naive theories in a domain suggest that certain features are "defining." We have already claimed that theoretical and conceptual knowledge are closely intertwined. Perhaps, then, the reason that people believe in a necessary basis for their concepts is that much of their knowledge of the world depends on correctly differentiating things into categories. Suggesting that concepts are ill-defined or fuzzy might cast doubt on much of one's knowledge.

However, not all features are perceived as defining; "defining" features, on our account, are those that are most central to our understanding of the world. In Fillmore's (1982) terms, those features that are most integrally involved with our idealized cognitive models will appear to be defining. For example, if it turned out that *carrots* weren't made of cells, then we would have to reconsider most of our other beliefs about carrots as well as about plants in general (for example, our theories of plant growth). Or if it turned out that some diamonds are really quite soft, then we would have to re-explain our past experiences with diamonds (or things we believed to be diamonds), the numerous claims people make about diamond hardness, our beliefs about diamond formation, and anything we might have known about crystal structures. Thus, being made of cells for carrots might be considered a defining feature, as might hardness for diamonds, because these features are so closely tied to other information about those categories.

If some of our characteristic features turned out to be wrong, a much smaller change in the knowledge base would be required. For example, if carrots weren't really orange, one could just assume they have been systematically dyed by unscrupulous grocers or farmers. This new information would probably not affect our concepts of plant life in general. If diamonds weren't really found in below-ground mines, none of the knowledge mentioned above would need to be reconsidered. One could assume that jewelers or diamond suppliers had lied in order to protect their market. In short, *defining features* are those at the meeting point of much of our knowledge.[8] *Characteristic features* are those toward the periphery of our knowledge base. More precisely, when a feature is involved in many causal links, rules, or scripts, it is perceived as "more defining" than a feature that is involved in few of them. The features at either end of the spectrum appear to be clearly defining or characteristic; those in the middle (involved in a moderate number of theoretical links) are the ones that cause arguments.

It is important here to separate the psychological question of defining and characteristic features from the philosophical–semantic issue. We think that, on reflection, most people would agree that it might be possible to find (or make) a soft diamond. Therefore, hardness is in some sense only a characteristic feature. Yet McNamara and Sternberg (1983) found that people say that being the hardest substance known is necessary for being a diamond. It seems likely that when people list such defining features, they are answering the question of which attributes are most central to their concepts, rather than which include all (potential) members and exclude all nonmembers. (An examination of other features given by McNamara & Sternberg's subjects reinforces this view.) Even if no feature is truly defining in a semantic–theoretical sense, people may put great weight on those that are tied up with much of their knowledge.

Conclusion

We have been arguing that people's theories and knowledge of the real world play a major role in conceptual coherence. This tendency to relate concepts and theories may be such that people impose more structure on concepts than simple similarity would seem to license.

Consider again the abominations of Leviticus, in which the animals that are clean and unclean for the people of Israel are listed in great detail. Over the years there have been many speculations concerning what properties of animals gave rise to their being listed as clean or unclean, as the overall similarity of the animals in each group is so low. To our minds, the most cogent speculation concerning this classification rule, developed in Mary Douglas's (1966) intriguing book, *Purity and Danger*, is that there should be a correlation between type of habitat, biological structure, and form of locomotion. Creatures of the water should have fins and scales, and swim; creatures of the land should have four legs and jump or walk; and creatures of the air should fly with feathered wings. Any class of

[8] Quine (1961) used a similar line of reasoning to argue against the existence of analytic truths, that is, statements that are necessarily true by virtue of the language. A prime candidate for such analytic truths has been to ascribe defining features to a concept, like "carrots are made of cells." Quine (1961, p. 43) pointed out that virtually *any* feature can be taken away from a category (e.g., hardness could be taken away from *diamonds*), but when some features are removed, a global reorganization of one's knowledge base is necessary. The larger this reorganization, the more analytic (defining) the feature is. Thus, he argued for a continuum of analytic to synthetic truths rather than a dichotomy. This philosophical argument parallels our psychological argument for why people perceive some features to be defining, although the two issues are potentially independent.

creature not equipped for the right kind of locomotion in its element is unclean. For example, ostriches would be unclean because they do not fly. Crocodiles are unclean because their front appendages look like hands, and yet they walk on all fours. If this analysis is correct, then there was a theory of appropriate physiological structure associated with each type of environment, and any animal that did not meet its standards was unclean. The category *clean animals*, then, comprises a coherent set of entities, even though the overall similarity of the members is very low. Although most categories probably have a better similarity structure than these examples, the point is clear that theories can impose coherence even when similarity is low.[9]

We think that there are two components to conceptual coherence. The first component involves the internal structure of a particular conceptual domain (see Table 2). Concepts that have their features connected by structure–function relationships or by causal schemata of one sort or another will be more coherent than those that do not. Although these correlations may be strictly empirical, in most cases they will be driven by expectations and hypotheses. In this way, the concept is integrated with the rest of the knowledge base. Other properties such as high within-category similarity and low between-category similarity may be by-products of this internal structure.

The second component of coherence has *to do with the position of the concept in the* complete knowledge base, rather than its internal structure (see Table 2). This component is the question of how the concept fits into "the cosmic machine revealed by science" (Quine, 1977, p. 171)—or, more accurately, the cosmic machine represented in people's heads. Concepts that have no interaction with the rest of the knowledge base will be unstable and probably soon forgotten. This component is also important in the formation of new concepts.

One objection to the theory-based approach that might be raised is that it is circular. How can mental theories explain concepts, the objection goes, when theories themselves are made out of concepts? The answer is that we are not attempting to *reduce* issues of conceptual representation to theoretical representation. On the contrary, we believe that the influence is bidirectional—one cannot talk about theories or knowledge representation in a domain without specifying the concepts people have in the domain. (In fact, research on people's naive theories has typically included discussion of their relevant concepts; see Gentner & Stevens, 1983.) Concepts and theories must live in harmony in the same mental space; they therefore constrain each other both in content and in representational format. Our point is that these constraints will provide insight into the structure of both areas, not that one can be replaced by the other. We agree that theories are made up of concepts (to a great extent) and urge that this fact be employed in our theories of concepts.

In our criticism of similarity as the sole basis of conceptual coherence, we pointed out that similarity needs to be greatly constrained before it makes any predictions. However, we should point out that the notion of a good theory is not yet fully constrained: We gave some idea in Tables 1 and 2 of what constitutes a good theory, but there is clearly more work to be done here. In fact, the point of this article is not to provide a complete account of the use of theories in conceptual structure, but is rather to demonstrate that theories are indeed important and to encourage future research to detail exactly how they are involved in concept formation and use.

We do *not* wish to suggest that previous *studies on novel concepts that are divorced* from real-world knowledge have no worth, nor that future such studies will be of little interest. These studies have provided the basis for our own theorizing, and they represent a necessary technique for studying conceptual structure. Our main point is that these studies and associated categorization theories relying exclusively on similarity relations are insufficient to provide a theory of concepts. We have argued that a coherent concept is one that we have a good theory about and that fits well with our other knowledge. This ap-

[9] We are guilty of oversimplifying here. No doubt the conceptual scheme associated with the division of clean and unclean animals is more elaborated and more intertwined with the culture that gave rise to these concepts than this example implies.

proach raises a number of empirical questions, many of them related to the question of how concepts are initially acquired and how expertise in a domain affects the concepts of that domain. The exact details of how theories affect internal and external conceptual structure have yet to be specified. Future research on concepts and categories can help answer these questions not by controlling the effects of world knowledge and experience, but by exploiting them—by bringing the concepts into contact with the whole cognitive system that created them.

References

Achinstein, P. (1968). *Concepts of science.* Baltimore: Johns Hopkins Press.

Adams, V. (1973). *An introduction to modern English word-formation.* London: Longman.

Adelman, L. (1981). The influence of formal, substantive, and contextual task properties on the relative effectiveness of different forms of feedback in multiple-cue probability learning tasks. *Organizational Behavior and Human Performance, 27,* 423–442.

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

Barsalou, L. W. (1982). Context-independent and context-dependent information in concepts. *Memory & Cognition, 10,* 82–93.

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition, 11,* 211–227.

Barsalou, L. W. (in press). Determinants of graded structure in categories *Journal of Experimental Psychology: Learning, Memory, & Cognition.*

Barsalou, L. W., & Bower, G. H. (1983). *A priori determinants of a concept's highly accessible information.* Unpublished manuscript, Emory University.

Berlin, B., Breedlove, D. E., & Raven, P. H. (1973). General principles of classification and nomenclature in folk biology. *American Anthropologist, 75,* 214–242.

Bower, G. H., & Masling, M. (1978). *Causal explanations as mediators for remembering correlations.* Unpublished manuscript, Stanford University, Psychology Department.

Bransford, J. D., Barclay, J. R., & Franks, J. J. (1972). Sentence meaning: A constructive versus interpretive approach. *Cognitive Psychology, 3,* 193–209.

Bulmer, R. (1967). Why is the cassowary not a bird? A problem of zoological taxonomy among the Karam of the New Guinea Highlands. *Man, 2,* 5–25.

Callanan, M. A., & Markman, E. M. (1982). Principles of organization in young children's natural language hierarchies. *Child Development, 53,* 1093–1101.

Camerer, C. F. (1981). *The validity and utility of expert judgment.* Unpublished doctoral dissertation, University of Chicago.

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, England: Cambridge University Press.

Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous diagnostic observations. *Journal of Abnormal Psychology, 72,* 193–204.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 74,* 272–280.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121–152.

Clark, E. V. (1973a). What's in a word? On the child's acquisition of semantics in his first language. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language* (pp. 65–110). New York: Academic Press.

Clark, E. V. (1973b). Non-linguistic strategies and the acquisition of word meanings. *Cognition, 2,* 161–182.

Clark, E. V. (1983). Meanings and concepts. In J. H. Flavell & E. M. Markman (Eds.), *Manual of child psychology: Cognitive development* (Vol. 3, pp. 787–840). New York: Wiley.

Clark, E. V., & Clark, H. H. (1979). When nouns surface as verbs. *Language, 55,* 767–811.

Clark, H. H. (1973). Space, time, semantics and the child. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language* (pp. 28–63). New York: Academic Press.

Clark, H. H., & Carlson, T. B. (1982). Speech acts and hearers' beliefs. In N. V. Smith (Ed.), *Mutual knowledge* (pp. 1–36). London: Academic Press.

Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. In J.-F. LeNy & W. Kintsch (Eds.), *Language and comprehension* (pp. 287–299). Amsterdam: North-Holland.

Cofer, C. N. (1973). Constructive processes in memory. *American Scientist, 61,* 537–543.

Cohen, B., & Murphy, G. L. (1984). Models of concepts. *Cognitive Science, 8,* 27–58.

Cohen, L. B., & Younger, B. A. (1983). Perceptual categorization in the infant. In E. K. Scholwick (Ed.), *New trends in conceptual representation: Challenges to Piaget's theory?* (pp. 197–220). Hillsdale, NJ: Erlbaum.

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences, 4,* 317–370.

Cole, M., & Scribner, S. (1974). *Culture and thought.* New York: Wiley.

Collins, A. (1978). Fragments of a theory of human plausible reasoning. In D. Waltz (Ed.), *Proceedings of the conference on Theoretical Issues in Natural Language Processing II* (pp. 194–201). Urbana: University of Illinois Press.

Collins, A., Brown, J. S., & Larkin, K. M. (1980). Inferences in text understanding. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 385–407). Hillsdale, NJ: Erlbaum.

Crocker, J. (1981). Judgment of covariation by social perceivers. *Psychological Bulletin, 90,* 272–292.

Day, J. C., & Bellezza, F. S. (1983). The relation between visual imagery mediators and recall. *Memory & Cognition, 11,* 251–257.

Dougherty, J. W. D. (1978). Salience and relativity in classification. *American Ethnologist, 5,* 66–80.

Douglas, M. (1966). *Purity and danger.* London: Routledge & Kegan Paul.

Fillmore, C. (1982). Towards a descriptive framework for spatial deixis. In R. J. Jarvella & W. Klein (Eds.), *Speech, place and action: Studies in deixis and related topics* (pp. 31–59). Chichester, England: Wiley.

Gelman, S. A., & Taylor, M. (1984). How two-year-old children interpret proper and common names for unfamiliar objects. *Child Development, 55,* 1535–1540.

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

Gentner, D., & Stevens, A. L. (Eds.). (1983). *Mental models.* Hillsdale, NJ: Erlbaum.

Geoghegan, W. H. (1976). Polytypy in folk biological taxonomies. *American Ethnologist, 3,* 469–480.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman, *Problems and projects* (pp. 437–447). Indianapolis: Bobbs-Merrill.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics, Vol. 3: Speech acts* (pp. 41–58). New York: Academic Press.

Hampton, J. A. (1981). An investigation of the nature of abstract concepts. *Memory & Cognition, 9,* 149–156.

Horton, M. S., & Markman, E. M. (1980). Developmental differences in the acquisition of basic and superordinate categories. *Child Development, 51,* 708–719.

Jenkins, J. J. (1974). Remember that old theory of memory? Well, forget it! *American Psychologist, 29,* 785–795.

Johnson-Laird, P. N. (1981). Mental models of meaning. In A. K. Joshi, B. L. Webber, & I. A. Sag (Eds.), *Elements of discourse understanding* (pp. 106–126). Cambridge, England: Cambridge University Press.

Jones, G. V. (1983). Identifying basic categories. *Psychological Bulletin, 94,* 423–428.

Karmiloff-Smith, A., & Inhelder, B. (1974/1975). If you want to get ahead, get a theory. *Cognition, 3,* 195–212.

Katz, N., Baker, E., & Macnamara, J. (1974). What's in a name? A study of how children learn common and proper names. *Child Development, 45,* 469–473.

Kay, P., & Zimmer, K. (1976). On the semantics of compounds and genitives in English. *Sixth California Linguistics Association Proceedings* (pp. 29–35). San Diego: Campile Press.

Keil, F. C. (1981). Constraints on knowledge and cognitive development. *Psychological Review, 88,* 197–227.

Keil, F. C., & Carroll, J. J. (1980). The child's conception of "tall": Implications for an alternative view of semantic development. *Papers and Reports on Child Language Development, 19,* 21–28.

Kelter, S., Grotzbach, H., Freiheit, R., Hohle, B., Wutzig, S., & Diesch, E. (1984). Object identification: The mental representation of physical and conceptual attributes. *Memory & Cognition, 12,* 123–133.

Kemler, D. G. (1982). Classification in young and retarded children: The primacy of overall similarity relations. *Child Development, 53,* 768–779.

Kuhn, T. S. (1962). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Lakoff, G. (1982). *Categories and cognitive models* (Cognitive Science Rep. No. 2). Berkeley: University of California, Cognitive Science Program.

Luria, A. R. (1976). *Cognitive development: Its cultural and social foundations.* Cambridge, MA: Harvard University Press.

Malt, B. C., & Smith, E. E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior, 23,* 250–269.

Markman, E. M., & Callanan, M. A. (1984). An analysis of hierarchical classification. In R. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 2, pp. 345–365). Hillsdale, NJ: Erlbaum.

McNamara, T. P., & Sternberg, R. (1983). Mental models of word meaning. *Journal of Verbal Learning and Verbal Behavior, 22,* 449–474.

Medin, D. L. (1983). Structural principles in categorization. In T. J. Tighe & B. E. Shepp (Eds.), *Perception, cognition, and development: Interactional analyses* (pp. 203–230). Hillsdale, NJ: Erlbaum.

Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 37–50.

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

Medin, D. L., & Schwanenflugel, P. L. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 355–368.

Medin, D. L., & Smith, E. E. (1984). Concepts and concept formation. *Annual Review of Psychology, 35,* 113–138.

Mervis, C. B., & Crisafi, M. A. (1982). Order of acquisition of subordinate, basic, and superordinate level categories. *Child Development, 53,* 258–266.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology, 32,* 89–115.

Michalski, R. S. (1983). A theory and methodology of inductive learning. *Artificial Intelligence, 20,* 111–161.

Michalski, R. S., Stepp, R. E., & Diday, E. (1981). A recent advance in data analysis: Clustering objects into classes characterized by conjunctive concepts. In L. Kanal & A. Rosenfeld (Eds.), *Progress in pattern recognition* (Vol. 1, pp. 33–55). Amsterdam: North-Holland.

Mill, J. S. (1965). *On the logic of the moral sciences.* New York: Bobbs-Merrill. (Originally published 1843)

Miller, G. A., & Johnson-Laird, P. N. (1976). *Language and perception.* Cambridge, MA: Harvard University Press.

Miller, P. McC. (1971). Do labels mislead? A multiple-cue study within the framework of Brunswik's probabilistic functionalism. *Organizational Behavior and Human Performance, 6,* 480–500.

Mohr, R. D. (1977). Family resemblance, platonism, universals. *Canadian Journal of Philosophy, 7,* 593–600.

Muchinsky, P. M., & Dudycha, A. L. (1974). The influence of a suppressor variable and labeled stimuli on multiple cue probability learning. *Organizational Behavior and Human Performance, 12,* 429–444.

Murphy, G. L. (1982a). Cue validity and levels of categorization. *Psychological Bulletin, 91,* 174–177.

Murphy, G. L. (1982b). *Note on measures of category structure.* Unpublished manuscript, Brown University, Psychology Department.

Murphy, G. L., & Smith, E. E. (1982). Basic-level superiority in picture categorization. *Journal of Verbal Learning and Verbal Behavior, 21,* 1-20.

Murphy, G. L., & Wisniewski, E. J. (1985). *Feature correlations in conceptual representations.* Unpublished manuscript, Brown University, Department of Psychology.

Murphy, G. L., & Wright, J. C. (1984). Changes in conceptual structure with expertise: Differences between real-world experts and novices. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 144-155.

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice-Hall.

Ortony, A., Vondruska, R. J., Jones, L. E., & Foss, M. A. (1984). Salience, similes, and the asymmetry of similarity. Unpublished manuscript, University of Illinois, Department of Psychology.

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. (1961). Two dogmas of empiricism. In W. V. O. Quine, *From a logical point of view* (2nd ed., pp. 20-46). New York: Harper & Row.

Quine, W. V. O. (1977). Natural kinds. In S. P. Schwartz (Ed.), *Naming, necessity, and natural kinds* (pp. 155-175). Ithaca, NY: Cornell University Press.

Richards, M. M. (1979). Sorting out what's in a word from what's not: Evaluating Clark's semantic features acquisition theory. *Journal of Experimental Child Psychology, 27,* 1-47.

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., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8,* 382-439.

Roth, E. M., & Shoben, E. J. (1983). The effect of context on the structure of categories. *Cognitive Psychology, 15,* 346-378.

Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 33-58). Hillsdale, NJ: Erlbaum.

Rumelhart, D. E. (1981). Understanding understanding (CHIP Rep. No. 100). San Diego: University of California, Center for Human Information Processing.

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

Sebetsyen, G. S. (1962). *Decision-making processes in pattern recognition.* New York: Macmillan.

Shweder, R. A., & Miller, J. G. (in press). The social construction of the person: How is it possible? In K. J. Gergen & K. Davis (Eds.), *The social construction of the person.* Berlin: Springer-Verlag.

Smith, E. E., Adams, N., & Schorr, D. (1978). Fact retrieval and the paradox of interference. *Cognitive Psychology, 10,* 438-464.

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

Suppe, F. (1977). The search for philosophic understanding of scientific theories. In F. Suppe (Ed.), *The structure of scientific theories* (2nd ed., pp. 3-241). Urbana: University of Illinois Press.

Tambiah, S. J. (1969). Animals are good to think and good to prohibit. *Ethnology, 8,* 424-459.

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

Tversky, A., & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology* (pp. 49-72). Hillsdale, NJ: Erlbaum.

Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General, 113,* 169-193.

Ullman, S. (1979). *The interpretation of visual motion.* Cambridge, MA: MIT Press.

Wattenmaker, W., Murphy, T. D., Dewey, G. I., Edelson, S. E., & Medin, D. L. (1984). Linear separability, knowledge structures, and concept naturalness. Unpublished manuscript, University of Illinois, Department of Psychology.

Wittgenstein, L. (1953). *Philosophical investigations.* New York: Macmillan.

Wright, J. C., & Murphy, G. L. (1984). The utility of theories in intuitive statistics: The robustness of theory-based judgments. *Journal of Experimental Psychology: General, 113,* 301-322.

Younger, B. A., & Cohen, L. B. (1984). Infant perception of correlations among attributes. *Child Development, 54,* 858-867.

Ziff, P. (1972). *Understanding understanding.* Ithaca, NY: Cornell University Press.

Received March 28, 1984
Revision received October 1, 1984 ■