Concepts and conceptual development: Ecological and intellectual factors in categorization

Edited by
ULRIC NEISSER

The right of the University of Cambridge to print and sell all manner of books was granted by Henry VIII in 1534. The University has printed and published continuously since 1584.

CAMBRIDGE UNIVERSITY PRESS
Cambridge
London New York New Rochelle
Melbourne Sydney

1987

3

Category cohesiveness, theories, and cognitive archeology

DOUGLAS L. MEDIN and
WILLIAM D. WATTENMAKER

Why do we have the categories we have and not others? The set of entities comprising our world could be partitioned in a virtually limitless variety of ways, but most of these partitionings prove to be vague, absurd, or useless, and only a tiny minority are informative, useful, and efficient. This chapter is concerned with the question of what makes categories psychologically cohesive.

The question of what makes a concept coherent has received a variety of answers, almost all of which rely directly or indirectly on the notion of similarity (similar entities appear in the same class, dissimilar entities belong to different classes). In the first part of this chapter we will argue that similarity-based approaches are inadequate, in part because the notion of similarity is too unconstrained and in part because these approaches fail to represent intra- and interconcept relations and more general world knowledge. Following Murphy and Medin (1985) we suggest, as an alternative approach, that concepts are coherent to the extent that they fit people's background knowledge or naive theories of the world.

The second part of our chapter is concerned with constraints on theories. If anything can qualify as a theory and all theories are equally good, then the problem of coherence simply has been shifted to a different level without being addressed. An obvious alternative to the idea that all theories are equally good is the idea that people prefer simple theories to more complex ones. We shall argue, however, that the notion of simplicity with respect to theories has many of the same limitations as the notion of similarity with respect to category structures. We then outline an alternative approach to constraints on theories that relies heavily on ecological considerations. The main idea is that the interaction of organisms with their environment provides a source of constraints that may become embodied in organisms. We refer to the search for such constraints as cognitive archeology. Although many of these constraints concern perceptual processing, they may be paralleled by analogous constraints in conceptual processing (e.g., theory construction). Although our particular line of argument is speculative, we believe that progress in analyzing constraints on both categories and the theories that are intertwined with them will necessarily reflect a sensitivity to the interaction of organisms and their environment.

Traditional approaches to conceptual coherence

For our present purposes it will prove convenient to work with a very informal definition of conceptual coherence. We use the term to describe groupings of entities that "make sense" to the observer as might be reflected in various measures such as ease of learning or even direct ratings of coherence. Coherence is not to be confused with the notion of naturalness as used by Keil (1981) or natural kinds as used by others because very unnatural concepts may also prove to be coherent in circumstances where the members of the category are coordinated through some theoretical frameworks. For example, consider the category comprised of the following objects: children, jewelry, portable television sets, photograph albums, manuscripts, oil paintings. Out of context such a category may not make much sense, but it becomes coherent in the context *things to take out of one's home during a fire.* Barsalou (1984; Chapter 5, this volume) has shown that these goal-derived categories behave very much like standard lexical concepts. Certainly these "ad hoc" categories are not "natural" by Keil's (1981) criteria, but they do seem to hang together in their own context. This example illustrates that coherence can be context-dependent. It is not even the case that taxonomic sorting is a cross-cultural universal. For example, Lancy (1983) has found that the Melpa of New Guinea use a pairing principle to create categories. That is, they attempt to specify a pair that, by way of contrast or complementarity between members of the two halves, forms a totality or whole. Melpa cosmology is not compatible with taxonomic categories.

The insufficiency of similarity

Perhaps the most intuitively plausible explanation of conceptual coherence is that objects, events, or entities form a concept because they are similar to each other. The basic idea is that similarity relations partition entities into natural clusters and that our concepts map onto these clusters. An immediate problem with this view is that entities may *seem* to be similar precisely because they are members of the same category. What is needed is some independent method of measuring similarity uncontaminated by people's knowledge about category memberships.

The predominant (and almost the sole) approach to being specific about similarity is to analyze concepts into constituent properties or attributes and to define similarity in terms of matching and mismatching properties. Thus, *robins* are more similar to *squirrels* than to *diamonds* because robins and squirrels share properties such as *living, mobile,* and *found in trees,* which are not possessed by diamonds.

Can the notion of similarity defined in terms of matching and mismatching properties explain why a category is formed (instead of some other) or its ease of use? As a paradigmatic case, consider Amos Tversky's (1977) theory of similarity, which defines it as a function of common and distinctive features weighted for salience or importance. An immediate problem with using Tversky's model to define category structure is that the similarity relationships among a set of entities will depend heavily on the particular weights given to individual properties or features. For example, a skunk and a zebra would be more similar than a horse and a zebra if the feature "striped" had sufficient weight. In some approaches to numerical taxonomy this issue is resolved by (somewhat arbitrarily) requiring that each feature be weighted equally. But Tversky (1977) has convincingly shown that 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 entity is to another (see also Gati & Tversky 1984). If this were not already a serious problem, Ortony, Vondruska, Jones, and Foss (1985) have argued that Tversky's model is too constrained in that it assumes that a given feature has the same salience regardless of the entity in which it inheres. If one is forced to the further concession that the salience of a particular feature can vary across entities (e.g., *stripes* for *barber poles* may be more salient than *stripes* for *bass* or some other species of fish), then it seems that there are too many free parameters and the notion of similarity becomes too flexible to explain coherence.

But perhaps the most serious problem with defining similarity in terms of common and distinctive attributes is that no constraints have been provided on what is to count as a feature or attribute. To illustrate this point Murphy and Medin (1985) argue that the number of attributes that *plums* and *lawnmowers* have in common could be infinite: both weigh less than 1,000 kilograms (and less than 1,001 kg), both are found in our solar system (on the earth, etc.), both cannot hear well, both have a smell, both can be dropped, both take up space, and so on. The list can be infinite. Any two entities can be arbitrarily similar or dissimilar depending on the criterion of what is to count as a relevant attribute.

At best, the notion of common and distinctive attributes provides a language for talking about similarity and representing conceptual coherence. 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 constrain which properties seem relevant and may make them seem similar.

It may seem that we are being a little harsh on such an important principle as similarity. We are not arguing that theories directly or dramatically alter appearances. Rather our claim is that the notion of similarity is too flexible to provide an account of conceptual coherence. The general form of argument we are advancing is that attempts to describe category coherence in terms of similarity will prove useful only to the extent that principles determining what is to count as a relevant property and determining the importance of particular properties are specified. In this case, however, it is important to realize that the explanatory work is being done by the principles that specify these constraints rather than the general notion of similarity. In many cases perceptual experience may seem to naturally partition entities into categories. One can think of the perceptual system as embodying a theory about what is important in the world. That is, the perceptual system has some built-in constraints on what will count as an attribute and what 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.

Even if one were to succeed in specifying perceptual constraints, the problem of conceptual coherence would not disappear. Some categorizations blantantly contradict perceptual similarity (e.g., categorizing whales as mammals) and the question of how much of our conceptual system is based on perceptually determined 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. We think that theories play a significant role in determining which properties are relevant but that the role of theories goes far beyond that. The notion that coherence derives from theories leads one to deemphasize individual attributes in favor of a focus on relational properties and the interaction of concepts in theorylike mental structures.

The insufficiency of correlated attributes

Although we have already argued that similarity does not sufficiently constrain concepts, it may be that some general processing principles that are based on similarity have greater explanatory power. One such principle is correlated attributes. Rosch and her associates (Rosch, Mervis, Gray, Johnson, & Boyes-Braem 1976; Rosch 1978) have proposed that natural categories divide the world up into clusters of correlated attributes that "cut the world at its joints." This clustering principle is in contrast to the idea that attributes occur in all possible combinations and are thereby uncorrelated. Natural object categories at the "basic level" (e.g., birds), which is neither the most specific nor most abstract level, are said to maximize the correlational structure of the environment by preserving these attribute clusters. In this view it is not undifferentiated "similarity" that makes concepts cohesive, but some more elaborated structure of correlations. An organism prepared 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.

Although the correlated attribute principle is attractive, it has several limitations. One minor problem is that a cause and an effect may be highly correlated, but they would probably be placed in different categories. A more serious problem is that, even with some predetermined set of properties, there are so many possible correlations that it is not clear how the correct ones get picked out (see Keil, 1981, for an extended discussion of this problem). It would seem that some additional principles are needed to further constrain category cohesion, such as the notion that correlations may be made more or less salient by people's theories. This latter point of view leads to a further concession, namely, that categories consistent with a theory, but violating correlational structure, may nonetheless be cohesive (Barsalou 1983, 1985).

The insufficiency of categorization theories

One might think that categorization theories might constrain similarity in such a way as to give an account of category cohesiveness. The major points of view concerning concepts, however, either say nothing about structure or take a syntactic approach that is so impoverished it ignores interproperty and interconcept relations.

Smith and Medin (1981) divide theories of category representation into three basic approaches: the classical view, the probabilistic view, and the examplar view. The classical view holds that all instances of a concept share common properties that are necessary and sufficient conditions for defining the concept. The probabilistic view denies that there necessarily are defining properties and instead argues that concepts are represented in terms of properties that are only characteristic or probable of class members. Membership in a category can thus be graded rather than all-or-none, where the better members have more characteristic properties than the poorer ones. The exemplar view agrees with the claim that concepts need not contain defining properties, but further claims that categories may be represented by their individual exemplars, and that assignment of a new instance to a category is determined by whether the instance is sufficiently similar to one or more of the category's known exemplars.

The classical view implies that the defining properties provide the structure that holds a category together. But this may not be enough. For example, a category consisting of purple things bigger than a basketball and weighing between 1.65 and 9.82 kilograms satisfies a classical view definition but does not seem sensible or cohesive. Osherson (1978) and Keil (1979) have worried about this problem and suggested that some of the needed constraints result from the hierarchical structure of ontological concepts that represent the basic categories of existence, such as *thing, physical object, event, solid,* and *fluid.* Although the status of ontological concepts is a matter under debate (e.g., Carey 1985; Gerard & Mandler 1983; Keil, Chapter 7, this volume), we think reinforcing the classical view with an ontological tree will represent a positive step toward developing constraints, at least to the extent that ontological concepts embody people's world knowledge. Of course, the remaining problem with the classical view as a structural principle is that many categories may not conform to the classical view (see Medin & Smith 1981; Mervis & Rosch 1981; Smith & Medin 1981, for reviews).

The probabilistic view is constrained primarily in that it implies that categories be partitionable on the basis of summing of evidence, that is, that the categories be perfectly separable on the basis of a weighted, additive combination of component information (this is called "linear separability" by Sebestyen [1962]). The probabilistic view has looser constraints in that categories conforming to the classical view are a proper subset of categories conforming to the property of linear separability.

The constraint of linear separability is important in certain algorithms for machine pattern recognition, but it does not seem to hold for people. One way of evaluating the importance of linear separability is to set up two categorization tasks that are similar in major respects, except that in one the categories are linearly separable, whereas in the other they are not. Using this strategy in a series of four experiments varying stimulus type, category size, and instructions, Medin and Schwanenflugel (1981) found no evidence that linearly separable categories were easier to learn than categories that were not linearly separable. Kemler-Nelson (1984) ran adults and children under conditions designed to induce either analytical or nonanalytical learning and also did not find that linear separability acted as a constraint.

The probabilistic view also inherits the problems associated with a simple syntactic approach – it provides no guidelines concerning which combinations of features form possible concepts and which form coherent ones. Thus, it would not rule out the following combination of typical features: bright red, flammable, eats mealworms, found in Lapland, and used for cleaning furniture. Clearly, the mere fact that this combination is probabilistic does not mean it is coherent.

Finally, the exemplar view provides no principled account of conceptual structure, since it does not constrain what exemplars are concept members. Although most exemplar theories assume that membership is based on similarity, we have already argued that this alone is not a satisfactory explanation of coherence.

The general insufficiency of attribute matching and similarity

We started this essay by noting that, without some constraints on what is to count as an attribute, similarity-based approaches to coherence relying directly or indirectly on attribute matching will not get off the ground. Actually, we believe that the problems with this approach go deeper, and that, in principle, they will prove to be insufficient. Some of the problems derive from the idea that conceptual structure can be understood in terms of a focus on constituent attributes, whereas others seem to follow from that more abstract principle that category membership is determined by a similarity-based matching process.

Focus on attributes. The practice of breaking concepts into constituent attributes engenders a tendency to view concepts as little more than the sum of their components. As Armstrong, Gleitman, and Gleitman (1983) noted, however, the simple fact is that most concepts are not a simple sum of independent features, whatever that be. For example, all the features that are characteristic of a bird do not make it a bird – unless these properties are held together in a "bird structure." This bird structure certainly consists of a large set of relational properties and not simply attributes.

One defense of the attribute-matching perspective is that relationships and operations 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 relationships mutually constrain one another. Consequently, whenever relational properties are present, attempts to represent component properties as independent will prove very awkward. Consider some artificial stimulus that contains a triangle inside a circle. One would need to present the presence of the triangle, the circle, and the *inside* relation the triangle bears to the circle. It will not work to represent the situation as a simple list (triangle, circle, inside) because that would not distinguish between a triangle inside a circle and a circle inside a triangle. Rather, one needs an internal structure that is more than a simple list and both structural constraints and explanatory power will derive from this richer structure. One might attempt to represent this complex internal structure in terms of a (higher-level) holistic attribute. To take this step, however, is to concede the inadequacy of the notion of independent properties and to abandon the goal of explaining similarity in terms of matching and mismatching attributes.

Categorization as attribute matching. A major respect in which attribute matching may be too limited is that our representations may include information concerning operations, transformation, and (indirectly) relations among properties. Consider the following example taken from Rips and Handle (1984). The participants in their experiment were asked whether an object 5 inches in diameter was more likely to be a coin or a pizza. Although the object's size was roughly midway between large coins and small pizzas (as determined by prior norms), participants tended to categorize it as a pizza. One reason for this judgment might be that pizzas are more variable in size and though the probability of a 5-inch pizza is low, it is higher than that of a 5-inch coin. As Rips and Handle (1984) point out, however, there may be more involved here than brute knowledge about variability. We know that coins may have a size mandated by law and this and related knowledge about the nature of coins supplies information about potential size and variability.

Again one might attempt to represent this knowledge in terms of complex attributes involving higher order properties. Although this may be technically correct, 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 consistent with these facts, it does not by itself explain or predict them.

Categorization as a matching process. Even if one could appropriately constrain attributes, relations, and the internal structure of attributes and relations, it still might prove misleading to think of categorization as involving a matching process establishing some form of identity between concept and exemplar. For example, the attributes and relations associated with higher level concepts and relations may be more abstract than those associated with lower level concepts or exemplars. Instead of matching of identities, 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 directly associated

with the concept *intoxicated,* yet that information might well be used to decide that a person is drunk. Categorizing the person as intoxicated may "explain" his or her behavior even though the specific behavior was not previously a component of the concept. Concepts may represent a form of "shorthand" for a more elaborate theory, and a concept may be invoked when it has sufficient explanatory relationship to an object rather than when it matches an object's attributes. That is, the relationship between a concept and examples may be closely analogous to the relationship between theory and data. If this is so, then it is misleading to represent categorization solely in terms of some form of matching.

Summary

We think that similarity-based approaches to category goodness are insufficient to explain conceptual coherence and the richness of conceptual structure. We choose the word *insufficient* because we do not wish to imply that this approach is completely wrong or misleading. Similarity, in an appropriately constrained form, may provide a natural explanation for at least some cases of conceptual coherence. One key to progress might rest with analyses of relational properties rather than a focus on attributes (we argue in the section on Theories, Relational Coding, and Linear Separability [below] that this is part of the solution). If one makes the plausible assumption that our perceptual apparatus becomes attuned (either ontogenetically or phylogenetically) to informative aspects of our environment, then an ecological approach to perception may provide some ways to pin down the notion of similarity. Even in this case it may prove useful to think of perceptual processes as embodying a theory about the world. At the same time it does not seem likely that one can draw a sharp distinction between perceptual and conceptual categories, and categories that seem grounded in perception may be significantly influenced by the sorts of conceptual structures developed on the basis of interaction with the associated entities. One can view our approach as supplying constraints missing from the (perceptual) similarity explanation, rather than simply contradicting it.

Table 3.1, taken from Murphy and Medin (1985), summarizes the differences between the similarity-based approach and the theory-based approach on a number of dimensions. We use "attribute" as a generic term for features, properties, propositions, and other independent chunks of knowledge, and "underlying principle" to refer to causal connects or explanatory relations that we have been describing as parts of theories.

One general way to characterize the difference between the two approaches is to say that the theory-based approach expands the bound-

Table 3.1. *Comparison of two approaches to concepts*

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

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

aries of conceptual representation. The theory-based approach implies that to characterize knowledge about a concept we must include a complex web of relations involving that concept and the other concepts that depend on it. At the very least, similarity-based approaches are going to need a much richer view of intra- and interconcept relationships than so far has been advanced, and the explanatory work is going to be done by a theorylike process that provides appropriate constraints.

The next section of this essay shifts focus from similarity-based to theory-based approaches to conceptual coherence. We review evidence that suggests concepts should be viewed as embedded in theories. This evidence will also serve to underline the limitations of similarity-based approaches to conceptual coherence.

The role of theories in conceptual coherence

Theories and attribute selection

For all we have said about the issue of what is to count as an attribute, it may be surprising that attempts to elicit attributes of concepts from experimental subjects have proven quite successful. The basic approach is quite simple: Ask participants to list properties of concepts and if several subjects list some attribute 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).

This attribute-listing technique has generated important data for categorization theories, but it raises the question of what determines which attributes are listed. Barsalou and Bower (1983) have made some progress on this issue in that they have shown that two types of properties are likely to be activated during processing. First, properties having high diagnosticity may be active because they are useful in distinguishing instances of a concept from instances of other concepts. Second, properties relevant to how people typically interact with instances of a concept are likely to be frequently active. These two principles do not exhaust people's conceptual knowledge. Thus properties that are necessary for category membership (e.g., for *birds*, having a heart, kidney, and lungs) might never be listed because they might not be discriminative properties in typical context and are only indirectly relevant to how people interact with them. These observations suggest that conceptual representations are not static entities and that attribute listing does not consist of some simple "readout" of attributes from some fixed structure. Rather, access is at least partially determined by typical forms of interaction, which may vary with context. In fact, Roth and Shoben (1983) have shown that typicality judgments vary as a function of particular contexts. For example, tea is a more typical beverage than milk in the context of librarians taking a break, but this ordering reverses in the context of truck drivers taking a break.

Actually, 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, or other specific birds. B. Tversky and Hemenway (1984) analyze this behavior in terms of cooperative rules of communication (Grice 1975) and implicit contrast sets (the diagnosticity principle – "two legged" does not distinguish robins from eagles). It appears then that attribute listings are constrained by a variety of factors other than simple truth conditions. People are unlikely to list *flammable* as a property of money, not because it does not burn but because flammability is not central to the role money plays in our theories of economic and social interaction. Therefore, attribute listing may predict goodness-of-example ratings because both attribute listings and goodness-of-example ratings are constrained by the same theories and systems of knowledge.

Theories and correlated attributes

We earlier considered the idea that coherent concepts are structured in terms of correlated attributes, but we expressed reservations because of computational problems associated with the many possible correlations. Theories may constrain correlated attributes in at least two distinct ways. First of all, theories may influence the salience of individual attributes and pairs of attributes. More fundamental, however, is the potential role of theories in linking properties in an explanatory system. For example, within the category bird, there is probably a correlation between the type of feather and whether or not the feet are webbed. But this is not a raw correlation that just happens to emerge, but rather a matter of logical and biological necessity. Adjusting to an aquatic environment may bring about a number of adaptations (e.g., webbed feet, water-repellent feathers) that would manifest themselves as correlated attributes. That is, people not only notice feature correlations, but they can deduce *reasons* for them based on their knowledge of the way the world works.

Not only is it the case that people can develop explanations for correlations, but also the availability of such explanations plays a causal role in the development of categories. We have recently completed a set of studies in which people were asked to sort descriptions into categories. In one case, for example, the descriptions were sets of symptoms and the categories were hypothetical diseases. The task was set up so that people could sort on the basis of a single property, or 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. Pilot work had suggested that some pairs of symptoms (e.g., dizziness and earaches, weight gain and high blood pressure) were easier to link than others (e.g., earaches to high blood pressure, dizziness to weight gain). People showed a strong tendency to cluster on the basis of correlated attributes for which a causal link could

be easily made. They also justified their sortings in terms of specific causal linkages explaining dizziness and earaches in terms of an ear infection that could disturb the vestibular organ. Thus feature correlations may be important in conceptual representations primarily when they can be represented as theoretical knowledge. The focus on correlated attributes underlines the importance of relational properties, and the study just described shows that theoretical considerations determine which relational properties are used to develop categories.

There is even evidence that theoretical expectations can dominate data in the perception of correlations. Chapman and Chapman (1967, 1969) presented evidence that therapists and undergraduate subjects using certain diagnostic tests perceived correlations between test results and psychological disorders when in fact there were none – or even when the opposite correlation occurred. For example, in the draw-a-person test, observers have the expectation that paranoia or suspiciousness of others will be revealed by how the eyes are drawn and these expectancies prevent observers from objectively evaluating this relationship. On the positive side, there is evidence that 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; Muchinsky & Dudycha 1974; Wright & Murphy 1984).

Theories, relational coding, and linear separability

Earlier we mentioned observations suggesting that linear separability is not a natural constraint on human categorization. Current categorization models implicitly assume that abstract structural constraints on categories hold across all realizations of that structure. We argue, however, that abstract category structures and knowledge structures interact to determine ease of learning. Linearly separable categories, for example, may be easier to learn than categories that are not linearly separable only in a limited number of contexts. The relative difficulty of the two structures may interact with the knowledge structures that are brought to bear on the task. Linear separability may not be an invariant or natural constraint because 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 important, linear separability would become important.

Some recent work by Wattenmaker, Dewey, T. Murphy, and Medin (1986) demonstrates this interaction of knowledge structures and abstract category structures. The motivation for this research derives from the argument that properties do not have the status of independent, irreducible primitives and consequently that the structure of interproperty relationships determines concept cohesiveness. For example, it is common practice to assume that the category *bird* is represented in terms of properties such as laying eggs, flying, having wings, building nests in trees, having feathers, and singing. Each of these components itself represents a complex concept with both internal structure and external structure based on interproperty relationships. Building nests is linked to laying eggs, and building them in trees poses logistic problems whose solution involves other properties (e.g., wings, hollow bones, flying). Given these complex relationships it is easy to see that there may be more to birdness than is captured by adding up a bunch of birdlike properties.

What type of interproperty encoding is compatible with using a summing strategy to learn linearly separable categories? Presumably, it is important that all features must be perceived to be related to some common theme. For example, consider apples, oranges, and pears as being analogous to component properties of an example. When the distinguishing characteristics of the components are salient, it makes little sense to sum the entities. If some superordinate concept becomes activated, however, that leads apples, oranges, and pears to be integrated (by being tied to the concept of fruit), then the notion of summing components may become sensible and linear separability may act as a constraint.

Exactly this logic was employed in the Wattenmaker et al. (1986) studies. In one experiment the examples used in learning consisted of sets of descriptions 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., made of metal, flat surface). The categories either were or were not linearly separable, and in one condition subjects were given the notion of hammer substitutes and in another condition they were not. The idea was that providing a theme would lead subjects to encode properties in terms of the superordinate theme – suitability as a hammer. If all the properties are encoded in relation to the hammer theme then a summing strategy should become natural and linearly separable categories will be easy to learn. The results showed a strong interaction of category structure with the presence or absence of the hammer theme. With the hammer theme the linearly separable categories were much easier to learn but in the absence of the theme the linearly separable categories actually proved harder to learn.

Other experiments by Wattenmaker et al. show that it is not the case that linear separability becomes important whenever the categories are meaningful. For example, in another study the examples were descriptions involving typical features of the occupational categories, *construction worker* and *house painter*. Here the additional clue was that the painter category consisted of both interior and exterior house painters. The categories which were not linearly separable contained correlated properties (e.g., between "works inside" and "works year round" and between "works outside" and "doesn't work in the winter"). The results of this experiment indicated that the additional theme facilitated the learning of the nonlinearly separable categories and impaired the learning of linearly separable categories.

These studies suggest that knowledge structures have systematic effects in terms of the forms of relational coding induced. The idea that linear separability is an important constraint can be thought of as holding for the special case where properties are encoded relative to an integrated category theme. One cannot describe some abstract category structure (e.g., the presence or absence of linear separability) as simple or complex, sensible or bizarre, independent of the form or theory and associated knowledge structure 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.

This work on theories and linear separability makes two major points. First of all, ease of learning or naturalness cannot be described simply in terms of matching and mismatching properties; rather, one needs to consider relational properties. Second, the salience of these relational properties is heavily constrained by the theory or theme intertwined with them.

Theories and prototype structure

The consensus view of the structure of natural object categories is that the overwhelming evidence showing typicality effects and the absence of evidence for defining features leads inevitably to the conclusion that categories are organized around characteristic or typical features. A popular theory of learning growing out of this view is that as a result of experience with examples of a category, people form an impression of the central tendency of a category and categorical judgments come to be based on this central tendency or prototype. Prototype theory, as an instance of the probabilistic view, shares all the limitations of the probabilistic view in terms of an account of category cohesiveness. This raises the question of how one might reconcile the ubiquitous typicality effects with concept naturalness. Again, we think the answer lies in viewing concepts as embedded in theories.

First of all, Barsalou (1985) has shown that typicality effects in goal-derived categories correlate highly with the degree to which examples satisfy the relevant goal or approximate the ideal value. He also performed similar analyses on common categories such as fruit and tools. Although the underlying goal or dimension for natural categories was 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 were partialed out. This suggests that natural concepts are at least partially organized in terms of underlying dimensions that reflect how the concept normally is involved with people's goals and activities.

Fillmore (1982) has made a related suggestion about the source of typicality structure. He argued that 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 a society in which certain (idealized) expectations about marriage and marriageable age are realized. The existence of "poor examples" of this concept – Catholic priests, homosexual men, men cohabiting with women friends, etc. – does not mean, Fillmore argues, that the concept itself is ill-defined. Rather, the claim is that the idealized cognitive model does not fit the actual world very well. Clearly, such a model is an example of what we have been calling "theories," since it provides a means of connecting many concepts to explain diverse facts. Mohr (1977) has argued that this is the correct way to view platonic universals, and Lakoff (1982; Chapter 4, this volume) has developed this notion of idealized mental models in some detail. For example, he has shown that typicality effects arise in part from interactions associated with multiple, overlapping cognitive models.

Fixed semantic structures and dynamic theories

Our focus on similarity and categorization theories may be a little misleading with respect to the arguments we have been making concerning interproperty and interconcept relations. Much of the research on semantic memory based on networks or sets of features can be viewed as attempts to represent just these relationships. Although attempts to represent these relationships in semantic structures allow for greater richness in conceptual representations, we see two fundamental problems with current approaches.

Johnson-Laird, Herrmann, and Chaffin (1984) argue that fixed semantic structures do not adequately capture people's knowledge about the world. To borrow one of their examples, people know that a tomato is more "squishable" than a potato, although it would seem implausible that this fact is directly represented in semantic memory. Furthermore, relational properties like "squishable" are highly dependent on context and the operations of freezing the tomato and boiling the potato would reverse the observations. It is very difficult to represent this flexibility of relationships in a fixed structure. The problem is not that relational properties are unconstrained but rather that much of our knowledge is computed rather than prestored. Network models do allow computation (e.g., inheritance links may allow one to infer that Helen of Troy had both arteries and veins), but it seems from examples like the operation of freezing that not all computations are wired into a network structure. It would seem that network models must be flexible enough to allow experiments to be run and evaluated "mentally" where the results of such experiments are constrained by world knowledge and people's theories about the world.

To us it seems that this pattern of flexibility and need for concern about extension to examples in the world leads naturally to viewing concepts as embedded in theories. Theories are dynamic and represent "mental models" of the world based on perception, memory, and imagination (see Johnson-Laird, 1983, for further arguments along these lines). Note that we are arguing for an ecological approach because the notion of mental models is almost vacuous without their extension to the world.

Ziff (1972) provides some delightful examples illustrating the role of mental models or conceptual schemes in understanding. For example, it seems sensible to say "a cheetah can outrun a man," but what about a one-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 circumstances, a cheetah would outrun a man. Ziff refers to this set of conditions as a conceptual scheme and makes the point that two people understand each other to the extent to which these conceptual schemes are shared. These implicit theories, which are likely quite flexible (imagine two veterinarians running a home for aged cheetahs) constrain our understanding of relationships among concepts.

Summary

We have been arguing that concepts should be viewed as embedded in theories. The idea is that coherence derives from both the internal structure of a conceptual domain and the position of the concept in the complete knowledge base (the external structure). Properties such as high within-category similarity and low between-category similarity may be by-products of internal structure as well as the rest of the knowledge base. The tendency to relate concepts and theories appears to be such that people impose more structure on concepts than simple similarity (if that can be defined) would seem to license.

Constraining theories

The arguments so far advanced do not constitute anything like a complete solution to the problem of conceptual coherence. Unless one can specify constraints on what a (good) theory is, it may not help to claim that conceptual coherence derives from having a theory. That is, the problem of category coherence may simply have been banished from the level of attributes and similarity only to reappear intact as the problem of theory coherence. This cannot be entirely true in that theories are characterized by an internal and external structure of causal linkages so that there is no clear analog to attribute matching. Still, one does not have to think this argument is devastating in order to take it seriously, and the remainder of this essay is directed toward attempts to specify constraints on theories.

The organization of this section is as follows. We begin with the notion that the principle of parsimony – the idea that simple theories are preferable to complex ones – is the only constraint needed. We contend that this approach is inadequate. We then turn to some of our recent work on rule induction, which further undermines simplicity as the sole constraint needed and which offers some guidelines for developing constraints on theories. This work is limited by exactly those factors that differentiate rules from theories, and we then return to a more ecological approach to constraints on theories. We conclude this section with a set of arguments that are responsible for the term *cognitive archeology* appearing in the title of this essay. The general need for an ecologically motivated approach to constraints on theories does not, of course, depend on the merits of these particular speculations.

Simplicity

The notion of simplicity or parsimony is so well engrained in the scientific community that one might wonder if other constraints are needed. Simplicity, however, is much like the concept of similarity in its elusiveness.

First of all, one needs to keep in mind that theories are related to data and the background for the argument about favoring the simpler of two theories is that they give an equally accurate account of the data. If the more complicated theory fits the data better, then already one is faced with the issue of trading accuracy for parsimony.

A second problem with the notion of simplicity is that it ignores the issue of the scope or domain of a theory. Both concepts and theories have an external structure defined in terms of a body of knowledge. Suppose one is comparing a theory that explains some observation in a simple way with an alternative theory that gives a more complicated account but which fits better with one's knowledge base more broadly construed. One might well prefer the latter theory. For example, one might, with some justification, opt for a very elaborate explanation of some experiment on extrasensory perception, because the alternative theory based on psychic powers does not fit in with a large body of other scientific theories and associated observations.

Perhaps the most serious problem with the notion of simplicity is that it depends upon the language of description employed. Consequently, simplicity can only be evaluated within a theoretical framework that is specified about what constitutes a basic operation and what is an elementary concept. Figure 3.1 illustrates this issue using a sequential induction problem. The task is to determine what the next figure (2 × 2 matrix) in the sequence ought to be. Presumably, the answer is that the most simple continuation is the one desired. The best continuation, however, depends on what is taken as basic. If one focuses on individual cells in the matrices and writes transition rules for successive figures (e.g., B – R, R – B for the top left cell, etc.), then the top fifth figure is clearly the most appropriate. If, instead, the complex 2 × 2 matrix is taken as basic, then one can write transition rules involving figure rotation (figures 3 and 4 are 90-degree counterclockwise rotations of figures 1 and 2, respectively) that lead one to the conclusion that the bottom fifth figure is the correct choice. If either choice were constrained to be described by a rule formulated in terms of the alternative basic operations, then the resulting rules would be very complex and unnatural. Again the point is that simplicity is not independent of the language of description.

A related problem with the notion of simplicity is that it is a product constraint (a simple theory) rather than a process constraint (theory generation). That is, it is consistent with the idea that people generate all sorts of theories varying in complexity and then select from those the

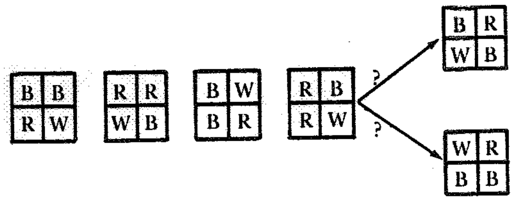

Figure 3.1. Which best fills in the fifth figure in the sequence?

most simple. Consequently, simplicity as a product constraint says little about how people come up with theories or how or why they might consider alternative descriptive languages. Although it is conceivable that simplicity constrains theories in just this way, we shall argue that it is more fruitful to look for constraints in terms of processes.

Rule induction and product versus processing constraints

There is a considerable literature on rule learning and much of it has been devoted to the relative difficulty of different types of rules (e.g., Haygood & Bourne 1965). This literature presents a mixed picture in that, for example, the greater ease of conjunctive than disjunctive rules may not be consistent over stimulus types (e.g., Reznick & Richman 1976). For our present purposes, we wish to emphasize that to the extent that research focuses on the relative difficulty of different types of rules at the expense of search for processing mechanisms, constraints will tend to be stated in terms of products rather than processes. Almost all of the work on rule learning has used experimenter-defined rules. In contrast to this approach, we have been evaluating candidate constraints using a rule-induction paradigm where subjects are presented with categories and asked to come up with a rule that not only will work for the set of examples but also could be used to classify new instances (Medin, Wattenmaker, & Michalski 1986). An example of this task is shown in Figure 3.2 where the requirement is to state a rule that could be used to decide if a train is East- or Westbound. The reader may wish to undertake this task before reading further.

It should be obvious that there is a large set of potential rules. Our goal is to develop principles that determine which subsets of these potential rules people find natural.

The task shown in Figure 3.2 is fairly difficult. Only a few people discovered the classifier using a simple property (East trains have three or more different loads). From a sample of 60 participants, we obtained 19 conjunctive rules (e.g., East trains have a triangle load and 3 or more loaded cars) and 32 disjunctive rules (e.g., West trains have two cars or a car with a jagged top). Given the background of research suggesting people have difficulty with negations, it is interesting that a fair number of rules involved negative properties (e.g., East trains have 3 or more cars and a triangle load and *not* a jagged top). These negative properties were almost exclusively associated with either conjunctive rules or the conjunctive part of a complex disjunctive rule (e.g., West trains have white engine wheels and not an oval-shaped car or else three circular loads). Therefore, it does not seem feasible to develop constraints at the level of products such as conjunctive versus disjunctive rules.

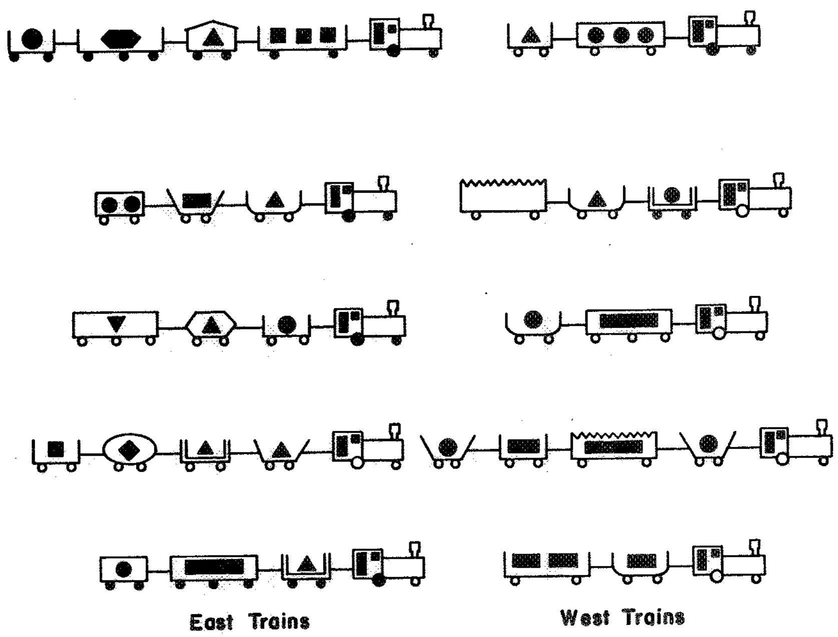

Figure 3.2. What separates the Eastbound trains from the Westbound trains?

What happens when one attempts to develop constraints in terms of processes? Without any pretense of doing more than providing an illustration, consider the following process model for the rule-induction task: People focus on one category and begin by looking for a descriptor that holds for all members of the positive (focus) set and does not apply to any counterexample. If one is found, then a simple rule can be generated. If no single descriptor works, because there are counterexamples, then one of two strategies may be applied. If there are numerous counterexamples, then people may look for combinations of properties (e.g., "X and Y") that span the set but do not generate counterexamples. If there are only a few counterexamples, then people may attempt to eliminate them by negating properties of the counterexamples not present in the positive set. For example, for the problem in Figure 3.2, a person may notice that all Eastbound trains have a triangle load but that two Westbound trains do also. This description is complete but not consistent. They might then look for combinations of properties that apply to the East but not the West trains. For example, they might consider the rule "triangle load in nonlast car," but that rule would still have a counterexample. Next a person might consider properties true of these two Westbound trains that are not shared by the East trains. For example, they might notice that the two West train counterexamples have a long car with two white wheels and then generate the rule "Eastbound trains have a triangle load and not long cars with two white wheels."

The other main possibility is that a descriptor has no counterexamples but fails to span the positive set. In that event people form a disjunction using the initial descriptor and then confine attention to the reduced positive set and the contrast set. For example, they might notice that only Westbound trains have two cars, and then focus on differences between the remaining two Westbound trains and the Eastbound trains. They might notice that the remaining West trains both have jagged tops and generate the rule "Westbound trains have two cars or a jagged top."

This account seems consistent with the present results. The descriptor, number of different loads, was apparently not very salient and few participants found the simple rule based on it. Number of cars was apparently quite salient and many people found the simple disjunction, number of cars and jagged top. According to this process model, negative descriptors (e.g., not jagged top) should be part of conjunctions and not part of disjunctions. As was mentioned earlier, this held for almost all cases where negative descriptors were used.

In this model the relative number of disjunctive and conjunctive rules would depend on the exact structure of the trains and the salience of the associated descriptors. In general, however, because people are assumed to focus initially on properties that members of the positive set have in common, conjunctive rules are likely to result. For the trains in Figure 3.3, the conjunction of two descriptors (e.g., dark wheels and closed top) has no counterexamples, and a simple conjunctive rule could be discovered. Exactly this preference for conjunctive rules was observed.

What about the results when the trains in Figure 3.3 were presented one at a time? In terms of our descriptive model the learning procedure should make it more difficult to discover properties that hold for all examples (i.e., are complete) or which do not have counterexamples (i.e., are consistent). If a person finds a property that is consistent but not complete (e.g., *short* for Westbound trains), the remaining train might be described in detail and one might see a rule like "West trains are *short* or *long with a circular load*." Note that such a rule is different from the rule "West trains are *short or have a circular load*" because it specifically combines circular load with long car. The disjunctive rules given in the sequential presentation condition conformed to the predicted pattern.

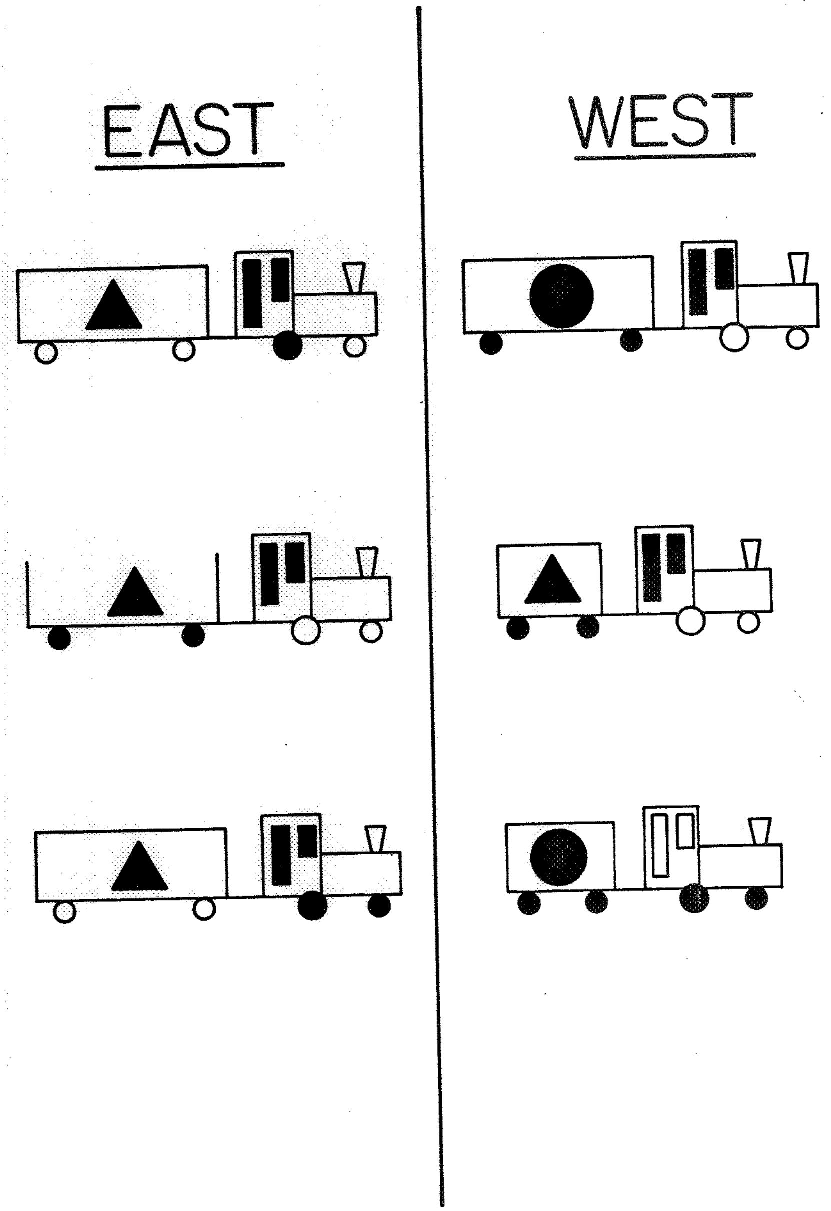

Figure 3.3. The less complex set of trains used as stimuli in the rule-induction paradigm.

The rule "short or a circular load" was never given, but rules of the form "*short* or *long with a circular load*" were fairly frequent.

Overall then, the results from the Medin, Wattenmaker, and Michalski rule-induction experiments are fairly consistent with the process model just outlined. The model could be formalized in terms of a computer program that embodied processing constraints and would be capable of inducing rules similar in form to those generated by human subjects. In fact, this series of studies had as its purpose to evaluate a specific inductive learning program, INDUCE (Michalski 1980, 1983a,b). Yet to be determined is the extent to which we are studying fairly general processing constraints as opposed to constraints associated with our particular tasks and stimulus materials. Although we do believe that the rule-induction studies underline the importance of developing processing constraints rather than product constraints, we are not under the illusion that theory construction and rule-induction are identical.

Rules versus theories. The work on rule induction just described has no semantics associated with it. The informal process model would say nothing about why, in our earlier study on correlated attributes, people preferred sorting by one pair of correlated attributes (those for which a causal connection could readily be made) rather than some alternative pair. Again, one needs to consider relational properties and the question of how theories might constrain them.

In one line of follow-up work we have used the trains shown in Figure 3.2 but presented different category labels and cover stories. For example, a participant might be told that the categories were trains run by smugglers versus legal trains, or trains constructed by creative versus uncreative children, or trains that travel in mountainous versus flat terrains. Our preliminary data suggest that different labels influence rule inductions in systematic ways. As one example of a change, the mountainous versus flat terrain labels make it much more likely that a participant will come up with the rule that the trains in one category have three or more different loads. When the smuggler category included the train carrying a diamond-shaped load, a few participants gave rules of the form "diamond-shaped load or . . ." even though the diamond descriptor applied to a single load. In addition, for these more meaningful categories, we have some evidence that participants are more likely to tolerate rules that are either incomplete or have counterexamples.

The follow-up work with meaningful labels suggests that one will need some form of representing people's real world knowledge in order to predict their rule inductions more accurately. The idea would be that this knowledge might make certain properties and combinations of properties more salient and thereby influence the particular rule inductions that people develop.

Actually, we think this example is quite deceptive in that it overlooks a crucial component of theory construction – the development of a descriptive language and basic operations associated with it. It simply is not the case that the world provides the basic units of analysis nor that competing theories necessarily subscribe to a common set of descriptive components. An illustration of this point is provided by a more recent rule-induction study (conducted in collaboration with Glenn Nakamura) using a richer set of stimulus materials. The category examples were children's drawings associated with the draw-a-person test used in clinical assessment. For a fixed set of drawings, we varied the category labels. For example, one set of participants might be told the drawings were done by mentally healthy versus disturbed children, another set told the drawings were done by creative versus noncreative children, and still another set told the drawings were done by farm versus city children. The results cannot be described simply in terms of the relative salience of a fixed set of properties, because category labels and descriptive units were not independent. For example, participants in one condition might note that the humans drawn by farm children all had at least some animal parts in them (e.g., a piglike nose), but when the same drawings were labeled as creative or mentally disturbed no participant mentioned the presence of animal parts.

One might be tempted to describe these results in terms of property salience changing across conditions, but we believe that to do so is to miss a very important point. This point is based on Kant's difference between the productive use of imagination (embodying innate perceptual constraints) and the reproductive use of imagination (involving relating particular experiences). Kant's terms (Kant 1933) correspond closely to Wittgenstein's distinction between "seeing the object" and "seeing the object as." The same object (e.g., a triangle) can be *seen as* a geometric drawing, a wedge, a mountain, a triangular hole, a threat, an arrow, and so on (see also Barresi 1981). We think this distinction is also important for the draw-a-person stimuli. Actually, in each of the different labeling conditions people stated rules at an abstract level and used the inferred properties of the drawings as support. For example, in the case of drawings supposedly done by farm children the rule might be that "each drawing reflects some aspect of farm life." One drawing might be seen to have a piglike nose, another to have a farmer's work clothes, and so on. These observations suggest that the drawings do not manifest some fixed set of properties that vary in salience so much as they "support" a limitless set of properties that derive from the interaction of the drawings with particular observers (Neisser develops this argument in Chapter 2, this volume).

Once one concedes that some set of stimuli or data "support" properties that hold only for the *interaction* of intelligent systems with aspects of their perceptual world, one must recognize that descriptive languages and basic operations may be different for different intelligent systems and may be different for the same individual at different points in time. In that sense, theory construction may represent a basic reorganization of knowledge rather than an increase or decrease in the salience of a set of prescribed properties. Susan Carey's (1985) studies of children's biological theories provide an elegant demonstration of a fundamental shift from a human-centered to a less egocentric organization of biological knowledge. Likewise, ether and phlogiston play no role in modern theories of physics. An adequate approach to constraints on theories cannot be forced to work with a fixed descriptive language because theories and descriptive languages are themselves mutually constraining. This is yet another reason why we do not see any hope for the idea that constraints can be developed in terms of products.

Naive theories

One straightforward approach to developing constraints on theories is to study people's theories about the world. There has been a recent upsurge of interest in people's mental models or naive theories (e.g., Gentner & Stevens 1983) of various physical phenomena. Many of these naive theories embody incorrect assumptions about the world and they often persist in the face of formal education. Consider the case of momentum. Suppose we have an airplane moving toward a target site to drop some cargo. Many people who are asked when the cargo should be released in order to land on the site say when the plane is directly above the site, failing to take the momentum of the cargo into account. From the point of view of looking for constraints on theories, these results are potentially of great interest in that they suggest that constraints associated with the learner lead to systematic inaccuracies.

So far, however, the work on naive theories has not led to major insights into constraints on theories. Perhaps one reason for this is that insufficient attention has been paid to an analysis of the structure of the environment in which naive theories are formulated. That is, we do not know the extent to which these naive theories are contradicted in experience. Few of us have the opportunity to drop things out of airplanes. A more typical circumstance involving momentum may be the situation of a person dropping something while in motion. But in this case, in trying to figure out when to release something, a person would also have to take into consideration their reaction time. The reaction time factor would lead them to initiate dropping of the object before they were directly over the target site. Appropriate calibration might come through experience, but because of the correlation between momentum and other factors (e.g., reaction time) there may be little in this experience that would lead a person to develop the notion of momentum for dropped objects. This line of reasoning is entirely speculative, but without an analysis of the environment in which naive theories are developed it hardly seems fair to call them naive, and at the very least it is difficult to use these observations to develop constraints on theories.

A closely related point is that insights into constraints on theories have been slow to come because greatest attention (perhaps quite appropriately, for many purposes) has been directed at the contents of naive theories rather than their structure. Typically, the naive theory is compared with the (more nearly) correct theory rather than the naive theory being compared with the environment in which the (naive) theory is formulated. It is possible that the structure of information in the environment is insufficient to allow one to converge on the correct theory (see Einhorn & Hogarth 1979). For example, psychiatrists do not see a random sample of the population and they frequently do not find out the fate of patients who do not return; anyone familiar with the principles of experimental design will recognize that these circumstances are far from optimal with respect to converging on an accurate impression of populations and treatments.

There is, of course, some work looking at the structure of naive theories. For example, the analysis of cultural belief systems in *The New Golden Bough* led to the principles of homeopathy (cause and effects tend to be similar) and contagion (a cause must have some form of contact to transmit its effect). Homeopathic medicine is based on the idea that a treatment must have physical resemblance to the symptoms associated with a disease. Despite the reservations we have expressed about the concept of similarity, some restricted form of similarity may act as a constraint or bias in theory formation. Incidentally, there is evidence that classical conditioning is more rapid if both the conditioned and unconditioned stimulus have a rapid onset or if both have a gradual onset than if one has a gradual and the other an abrupt onset (Testa 1974). Again however, we add that, without further analyses, it is difficult to know the extent to which homeopathy is supported by the environment.

There is also evidence that analogies and metaphors rather than just literal similarity serve to structure theories. In an analysis that parallels our reservations about attribute matching, Gentner (1983) has argued that analogies and metaphors involve a translation of relational structures rather than a mapping of attributes or properties. This line of work makes the important point that theories often are constrained by the fact that theories in one domain often are exported to another domain (for some interesting examples taken from psychology, see Gentner & Grudin 1985). Consider, for example, theories about extroversion and introversion. Based on a reservoir metaphor, one views extroverts as having excess energy that is often expressed socially. On the other hand, if one takes homeostasis as the appropriate metaphor, then one might think of extroverts as being chronically underaroused and seeking social stimulation. This analysis of analogy and metaphor strikes us as a very promising approach to constraints, and it should pay off to the extent that principles associated with mapping from one domain into others can be developed (see Gentner 1983). Of course, this transference from domain to domain will need to be traced back to some original theory or set of theories.

Cognitive archeology

By now it should be clear, if not by elegance of presentation then certainly by repetition, that we think that constraints on theories should be developed in terms of a process model growing out of an analysis of the relationship between people and their environment. But one needs a strategy for going about this. In this section we describe a strategy based on the work of Gibson, Neisser, and Shepard, which we refer to as cognitive archeology, and then suggest a heuristic device derived from this strategy for identifying constraints on theories.

Relationships between organisms and the environment. One of Gibson's major contributions was to argue that perception is active rather than contemplative and passive, and that the perceptual system represents an adaptation of organisms to their environment (e.g., Gibson 1979). The environment provides an anchor and perceptual systems are constrained to respond to the information embodied in that environment. For example, according to Gibson, *constancy* is not a mental construct but rather an objective property of the environment.

In his recent discussion of Gibson's work, Russell (1984) focuses on the adaptive function of perception. In his words, "So one way of describing the message of TDP [Theory of Direct Perception] is that we should not be overimpressed by the analogy between a perceptual system and a cognitive system, but should bear in mind that in the other direction there is an analogy with a respiratory system" (Russell 1984:167). One could conceive of breathing as the perception of oxygen but that would probably yield few insights into the process of respiration. Analyzing perception in terms of cognitive operations, Russell argues, may be similarly unilluminating.

Gibson recognized that there was more to perception than knowledge of properties of objects. There are also *affordances* which are based on *relationships* between objects and particular organisms. Every object is involved in infinitely many objective affordances, only some of which are specified unambiguously by optical information. Neisser (Chapter 2, this volume) extends this view by including the notion of conceptual properties, which are defined by the relationship between objects and certain systems of cultural or scientific beliefs. Thus perception is direct, whereas categorization is not, because categorization depends on the fit between objects and the theories we make about them.

Aside from the general point that theories are about the world, Neisser's view links theories to the perception of the properties of objects and affordances associated with the interaction of organisms and their environment. One might wonder whether these linkages or moorings are sufficient to keep the ship of conceptual properties and theories from going adrift. The main point we wish to make, however, is that organisms reflect the (evolutionary) history of their interaction with their environment in terms of adaptive specializations, and this is as true for the cognitive system as it is for the respiratory system.

Adaptive specializations can take a variety of forms. Consider, for example, the case of nest building in the long-tailed tailorbird, which actually sews leaves together to provide a superstructure for its nest. Conceivably, tailorbirds are very intelligent and nest building reflects just one instance of elaborate forms of learning that take place in each generation. Another possibility is that tailorbirds have a specialized ability to learn how to build nests that does not extend to other domains. Of course, the entire pattern of behavior could be innate and, in fact, for nest building in tailorbirds, that is the case. Since adaptive specializations that attune organisms with the demands of their environment can be expressed in varied forms, it may take experimental intervention to determine the particular form of adaptation that is present. For example, many organisms display day – night cycles or biological (circadian) rhythms in their behavior. These cycles could be totally driven by the presence or absence of light. When rats are placed in continuous darkness, however, they continue to manifest cycles of activity that closely mirror day and night. This result suggests that rats have some form of "internal clock." Although this clock is not accurate to the second and exposure to prolonged periods of darkness can lead to asynchrony with day and night, these clocks quickly become calibrated when rats are reexposed to a normal pattern of day and night. In brief, the form of adaptation associated with circadian rhythms in rats was discovered by breaking the normal correlation between the environment and the organism.

The above example is the key to a research strategy outlined by Roger Shepard in a recent essay (Shepard 1984). Shepard argues that an ecological approach to perception does not imply that one should only study perception in ecologically valid contexts. To do so would raise the problem of separating constraints associated with the environment itself from those embodied in the organism. To discover which constraints are embodied, Shepard suggests that one needs to provide an ambiguous or neutral context and see what comes out of this decoupling of organism and typical environment. Note that this is not a license to use any experimental situation; rather the situation has to be carefully structured to break this organism by environment correlation. The benefits of this strategy are nicely documented by Shepard's (1984) review of his program of research on perception.

At this point the reader may be wondering what this work on perception has to do with constraints on theories. In the second half of this essay we have been arguing that an ecological approach to constraints on theories is appropriate. We now wish to suggest, as a heuristic device, that constraints on theories may show strong parallels to embodied constraints associated with perception as well as with learning and memory. Before elaborating on this idea we note that it is not particularly novel – Lakoff and others have noted, for example, that many of our most important metaphors are based on our understanding of space and spatial relationships (Lakoff & Johnson 1980).

Cognitive archeology. The notion of cognitive archeology is directly based on Shepard's research strategy and the idea that constraints associated with the environment are embodied or reflected in organisms. Just as one might reason from properties in some world what types of organisms could live in it, so also one might reason from properties of some organism something about the world (environmental demands) they were part of. Decoupling provides an opportunity to evaluate the form in which constraints or biases have been embodied as opposed to being supported by the environment alone (see Birnbaum, 1975, for an interesting methodological approach to decoupling). The Chapmans' work on illusory correlation is an example of this strategy in that it shows that people develop theories about relationships between responses on projective tests and diagnostic categories that are not objectively supportable (Chapman & Chapman 1967, 1969). Since these incorrect theories tend to be shared theories (observers agree with each other), it seems likely that some basic constraints on theory formation are reflected in this phenomenon.

Consistent with the notion that constraints on theories may show strong parallels to embodied constraints associated with more basic processes. Rozin (1976) has argued that evolution involves a freeing up of special purpose mechanisms and that systems which are initially "tightly wired" and unconscious may ultimately give rise to consciousness and flexibility. As one example of special purpose mechanisms, many of the inferential systems involved in visual perception are tightly wired into the visual system and inaccessible to consciousness (see Ullman, 1979, for unconscious computational processes associated with the perception of motion). The essence of Rozin's argument is that evolutionary mechanisms result in the application of adaptive specialization from one domain to other domains. Thus, special purpose mechanisms originally tightly wired into basic processes (e.g., visual perception) may gradually become accessible to other systems (e.g., cognitive processes) and, in the extreme, may give rise to consciousness and flexibility.

If the evolution of cognitive systems is associated with a gradual increase in consciousness and flexibility, then there may be some close relationships among the mechanisms associated with basic learning, memorial, and perceptual processes and higher cognitive processes such as theory building. That is, we are suggesting that conceptually based theories may preserve some of the constraints or biases associated with more basic (less flexible) cognitive systems. Although we think that Rozin's ideas on conscious access are intriguing, our suggestion does not hinge on the issue of consciousness. Whether or not there is a close connection between perceptual and conceptual constraints is completely independent of the awareness question.

Our suggestion, then, is that constraints associated with what we normally think of as more basic (and less accessible) cognitive processes may carry over into higher-level cognitive processes. That is, a process model for theory construction may embody many of the same constraints as embodied in process models for learning, memory, and perception. It is not easy to come up with illustrations of this point, but we shall offer an example. In the section on Naive Theories we did not cite any of the vast literature on people's ordinary explanations and causal attributions associated with observing themselves and other people. We find it intriguing, however, that the principles being developed to understand people's causal attributions show some striking similarities to principles embodied in modern theories of animal conditioning (for examples of the former see chapters by Fincham and by Kelley in the volume edited by Jaspars, Fincham, & Hewstone 1983; for examples of the latter, see Rescorla & Wagner 1977, and Rescorla & Durlach 1981). For example, both domains rely on a principle of informativenss (e.g., in conditioning, learning is thought to occur only to the extent that the unconditioned stimulus is not expected) and both assume that temporal contiguity, spatial proximity, and similarity of magnitudes (of cause and effect or of conditioned and unconditioned stimulus) influence performance (attribution and conditioning, respectively). The parallels are close enough that people in animal conditioning are taking attribution theory seriously (e.g., Bolles 1976; Dickinson, Shanks, & Evenden 1984) and examination of the literature on animal conditioning might prove to be a useful source for identifying candidate constraints on causal attributions.

There may be even closer ties between perception and causal attribution. Michotte's (1963) classic work on the perception of causality still makes fascinating reading, and he himself was well aware of the potential link between causal perception and thought. In his words, "There are some cases, viz. launching and entraining, in which a causal impression arises, clear, genuine, and unmistakable and the idea of cause can be derived from it by simple abstraction in the same way as the idea of shape or movement can be derived from the perception of shape or movement" (Michotte 1963:270–271).

For a second example of our point, we turn once again to correlated attributes. Rosch, Mervis, and their associates (see Mervis & Rosch, 1981, for a review) have argued that the world is structured in terms of clusters of correlated attributes. Sensitivity to correlated attributes might be embodied in a (large) computational device that stores information about pairs of attributes, triples, and so on (see Hayes-Roth & Hayes-Roth 1977). Aside from the problem mentioned in the first section of this paper concerning the many possible correlations, we think something as important as sensitivity to correlated attributes would likely not be manifest in a general purpose computational device. As one alternative, the Medin and Schaffer (1978) context model assumes that people store examples of concepts and that categorization decisions are based on reasoning by analogy from examples that are retrieved by the probe. For instance, some novel four-legged creature might be classified as a mammal because it is quite similar to an elephant (or in another case to a mouse) and the categorizer may know that elephants (or mice) are mammals. The context model constrains similarity to be an interactive function of matching and mismatching properties. Although the model is noncomputational and makes no direct assumptions about the encoding of attribute correlations, it nonetheless behaves in a manner that is sensitive to correlated attributes (e.g., Medin, Altom, Edelson, & Freko 1982). Incidentally, there is evidence that people's categorization decisions reflect a sensitivity to correlated attributes even when they are totally unable to verbalize the basis of their decisions (Lewicki 1985; Wattenmaker 1986). People's sensitivity to correlated attributes appears to have both analytic and nonanalytic components.

The general point is that there may be interesting linkages among structural constraints in the environment, nonconscious and relatively inflexible perceptual and memorial mechanisms, and more flexible theories. As an aside, the context model evolved out of analyses of animal learning and memory (see Medin & Reynolds, 1985, for a review). Regardless of whether the context model is otherwise a good or bad model, it seems likely that we will need categorization models that can achieve sensitivity to correlated attributes without conscious computations. Furthermore, it seems possible that more conscious, flexible theorizing may inherit or not deviate substantially from these constraints. For example, Medin and Smith (1981) asked different groups of subjects to employ different strategies in the same categorization task. These instructions produced large phenotypic differences in performance, but in each case the context model was able to provide an accurate fit to the data by the assumption that strategies worked to change the salience of different properties comprising the examples.

To extend this example a bit, we have also recently examined a modified version of the context model where it is assumed that if responding by analogy with retrieved information is successful, then similarities between the current material and what was retrieved are made more salient, whereas if the analogy fails differences between the current material and what was retrieved are made more salient. These assumptions lead the model to behave in a rulelike way without directly incorporating rules, just as earlier versions of the model led to sensitivity to correlated attributes without computing correlations. We refer to the representations that give rise to rulelike performance as rule precursors. What is interesting about these rule precursors is that they can be used to predict the explicit rules that people give in our rule-induction task. For example, for the trains shown in Figure 3.3, simulations of this modified context model lead to precursors that behave like a conjunctive rule or like a rule plus exception strategy, but we do not observe precursors that behave like a simple disjunction. In other words, the constraints associated with these precursors may tend to be paralleled by corresponding constraints on the explicit rules that people give us.

It may be that the above parallels between noncomputational processes and rules or strategies either are accidental or occur for some trivial reason. We have only begun to study nonanalytic learning to see if people's rulelike behavior conforms to predictions of the context model, and it should be clear that our example is intended to be only illustrative. We do think it is important to realize that the mere fact that theories are (usually) about the world may not sufficiently constrain them. Our speculation is that flexible, conscious (and computational) theorizing is constrained in terms of processing principles which may not stray too far from less flexible processing mechanisms that embody constraints associated with the interaction of intelligent organisms with their world. The apple may not fall far from the tree.

Summary

In this chapter we have argued that similarity-based approaches to conceptual coherence are insufficient to explain the richness of conceptual structure, and opt instead for a theory-based approach to conceptual coherence. A theory-based approach emphasizes that coherence derives from both the internal structure of a conceptual domain and the position of the concept in the complete knowledge base. Concepts are viewed as embedded in theories and are coherent to the extent that they fit people's background knowledge or naive theories about the world.

The second half of the chapter discussed possible strategies for identifying constraints on theory formation. Theories were viewed as being extensionally anchored by the interaction of people with their world both in terms of the structure manifest in the world and in terms of the structure embodied in the human organism. It was emphasized that the search for constraints should be governed by an attempt to develop process models of the interaction between people and their environment. We speculatively suggested that a useful strategy for developing such models might be to examine the possibility that constraints on theories show strong parallels to embodied constraints associated with basic perceptual, learning, and memorial processes. Adaptive specializations to the environment that are initially tightly wired into a particular system and accomplish specific purposes may eventually be incorporated into other systems, and may ultimately give rise to consciousness and flexibility. If the evolution of cognitive systems is associated with a gradual increase in accessibility, then there may be some close relationships among the mechanisms associated with basic perceptual and cognitive processes and higher cognitive processes such as theory construction. Conceptually based theories may preserve some of the constraints or biases associated with more basic and less flexible cognitive systems.

NOTE

Preparation of this paper was supported by National Institute of Mental Health grant MH32370 to Douglas L. Medin. We would like to thank Dedre Getner, David Green, and Sarah Hampson for their helpful comments on an earlier draft of this manuscript.

REFERENCES

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.

Baressi, J. (1981). *Perception and imagination.* Paper presented at the Conference on the Philosophy of Perception and Psychology, Montreal, Canada.

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

(1984). *Determinants of graded structure in categories.* Unpublished manuscript, Emory University Psychology Department, Atlanta, GA.

(1985). Ideals, central tendency, and freqeuncy of instantiation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 629–654.

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

Birnbaum, M. H. (1975). Expectancy and judgment. In F. Restle, R. Shiffrin, W. G. Castellan, H. Lindman, & D. Pisoni (Eds.), *Cognitive theory* (Vol. 1). Hillsdale, NJ: Erlbaum.

Bolles, R. C. (1976). Animal learning and memory. In D. L. Medin, T. J. Roberts, & R. M. Davis (Eds.), *Animal memory.* Hillsdale, NJ: Erlbaum.

Carey, S. (1985). *Conceptual change in childhood.* Cambridge, MA: M.I.T. Press.

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

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

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, IL: University of Illinois Press.

Dickinson, A., Shanks, D., & Evenden, J. (1984). Judgment of act-outcome contingency: The role of selective attribution. *Quarterly Journal of Experimental Psychology, 36A*(1), 29–50.

Einhorn, H. J., & Hogarth, R. M. (1979). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review, 85,* 395–416.

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.

Gati, I., & Tversky, A. (1984). Weighting common and distinctive features in perceptual and conceptual judgments. *Cognitive Psychology, 16,* 341–370.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7,* 115–170.

Gentner, D., & Grudin, G. (1985). The evolution of mental metaphors in psychology: A ninety-year retrospective. *American Psychologist, 40*(2), 181–192.

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

Gerard, A. B., & Mandler, J. M. (1983). Ontological knowledge and sentence anomaly. *Journal of Verbal Learning and Verbal Behavior, 22,* 105–120.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston: Houghton Mifflin.

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.

Hayes-Roth, B., & Hayes-Roth, F. (1977). Concept learning and the recognition and classification of exemplars. *Journal of Verbal Learning and Verbal Behavior, 16*, 321–338.

Haygood, R. C., & Bourne, L. E., Jr. (1965). Attribute and rule-learning aspects of conceptual behavior. *Psychological Review, 72*, 175–195.

Jaspars, J., Fincham, F., & Hewstone, M. (Eds.) (1983). *Attribution theory and research: Conceptual, developmental and social dimensions.* London: Academic Press.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness.* Cambridge, England: Cambridge University Press.

Johnson-Laird, P. N., Herrmann, D. J., & Chaffin, R. (1984). Only connections: A critique of semantic networks. *Psychological Bulletin, 96*(2), 292–315.

Kant, I. (1788/1933). *Critique of Pure Reason.* Translated by Norman Kemp Smith. London: Macmillan.

Keil, F. C. (1979). *Semantic and conceptual development: An ontological perspective.* Cambridge, MA: Harvard University Press.

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

Kemler-Nelson, D. G. (1984). The effect of intention on what concepts are acquired. *Journal of Verbal Learning and Verbal Behavior, 23*, 734–759.

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

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by.* Chicago: University of Chicago Press.

Lancy, D. C. (1983). *Cross-cultural studies in cognition and mathematics.* New York: Academic Press.

Lewicki, P. (1985). Processing covariations among features. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*, 135–146.

Medin, D. L., & Reynolds, T. J. (1985). Cue-context interactions in discrimination, categorization, and memory. In P. Balsam & A. Tomie (Eds.), *Context in learning and memory*, Hillsdale, NJ: Erlbaum.

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. (1981). Strategies in classification learning. *Journal of Experimental Psychology: Human Learning and Memory, 7*, 241–253.

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., Wattenmaker, W. D., & Michalski, R. S. (1986). *Constraints in inductive learning: An experimental study comparing human and machine performance.* Manuscript submitted for publication.

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

Michalski, R. B. (1980). Pattern recognition as rule-guided induction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(4), 349–361.

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

(1983b). A theory and methodology of inductive learning. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning* (pp. 83–134). Palo Alto, CA: Tioga Publishing.

Michotte, A. (1963). *Perception of Causality.* London, Methuen.

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

Ortony, A., Vondruska, R. J., Jones, L. E., & Foss, M. A. (1985). Salience, similes, and the asymmetry of similarity. *Journal of Memory and Language, 24*, 569–594.

Osherson, D. N. (1978). Three conditions on conceptual naturalness. *Cognition, 6*, 263–289.

Rescorla, R. A., & Durlach, P. J. (1981). Within event learning in Pavlovian conditioning. In N. E. Spear & R. R. Miller (Eds.), *Information processing in animals: Memory mechanisms* (pp. 81–112). Hillsdale, NJ: Erlbaum.

Rescorla, R. A., & Wagner, A. R. (1977). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II* (pp. 64–99). New York: Appleton-Century-Crofts.

Reznick, J. S., & Richman, C. L. (1976). Effects of class complexity, class frequency, and pre-experimental bias on rule learning. *Journal of Experimental Psychology: Human Learning and Memory, 2*, 774–782.

Rips, L. J., & Handle, J. (1984). *Classification without similarity.* Unpublished manuscript, University of Chicago.

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. G. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573–605.

Rosch, E., Mervis, C. G., 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.

Rozin, P. (1976). The evolution of intelligence and access to the cognitive unconscious. In J. M. Sprague & A. N. Epstein (Eds.), *Progress in psychobiology and physiological psychology* (pp. 245–280). New York: Academic Press.

Russell, J. (1984). *Explaining mental life.* London: Macmillan Press.

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

Shepard, R. N. (1984). Ecological constraints on internal representation: Resonant kinematics of perceiving, imagining, thinking, and dreaming. *Psychological Review, 91*(4), 417–447.

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

Testa, T. J. (1974). Causal relationships and the acquisition of avoidance responses. *Psychological Review, 81*(6), 491–505.

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

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. D. (1986). Nonanalytic concept formation and sensitivity to correlated attributes. Manuscript in preparation, University of Illinois at Champaign-Urbana.

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

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.

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