COGNITIVE PSYCHOLOGY 20, 158–190 (1988)

Context and Structure in Conceptual Combination

DOUGLAS L. MEDIN AND EDWARD J. SHOBEN

University of Illinois

Three experiments evaluated modifications of conceptual knowledge associated with judgments of adjective–noun conceptual combinations. Existing models, such as the Smith and Osherson modification model, assume that the changes associated with understanding an adjective noun combination are confined to the corresponding adjectival dimension. Our experiments indicate that this assumption is too strong. The first study found that naming one dimension affects correlated dimensions. For example, participants judge small spoons to be more typical spoons than large spoons, but for wooden spoons, large spoons are more typical than small spoons. The second study demonstrated that the similarity of adjectives is not independent of the noun context in which they appear. For example, white and gray are judged to be more similar than gray and black in the context of *hair* but this judgment reverses in the context of *clouds*. The third study showed that a property equally true (or false) for two concepts may be more central to one concept than the other (e.g., it is more important that boomerangs be curved than that bananas be curved). These results pose serious problems for current theories of how people combine concepts. We propose instead that we need richer views of both the conceptual structure and the modifications of it required by conceptual combination. We suggest that theoretical knowledge and the construct of centrality of meaning may play useful roles.

Many attempts have been made to characterize the structure of conceptual knowledge in human behavior. Not unexpectedly, first attempts have assumed that knowledge is comprised of relatively static and well-bounded packets of information (e.g., definitions, prototypes, schemata, frames, scripts), each of which represents some particular kind of object or event (e.g., whale, eating in a restaurant, birthday party). Furthermore, implicit in many accounts of conceptual structure is the assumption that the same packet of information is employed in a wide range of contexts, including the case where a concept is used in combination with other concepts (as when *birthday* and *party* are combined into *birthday party*). For example, when people are asked to list examples of a concept such as bird, typical examples such as robin and sparrow are more likely

This research was supported in part by NSF Grant BNS84-19756 and National Library of Medicine Grant LM04375 to the first author and by NSF Grants BNS82-17674 and BNS86-08215 to the second author. Kenneth Gray provided valuable assistance in the conduct of the first experiment and preliminary versions of the third experiment. We thank Marie Banich, Don Dulany, and Brian Ross for helpful comments on earlier drafts of this manuscript. Address correspondence, including requests for reprints, to either Douglas L. Medin or Edward J. Shoben, Department of Psychology, University of Illinois, 603 E. Daniel, Champaign, IL 61820.

0010-0285/88 $7.50

to be mentioned and are mentioned earlier than atypical examples such as peacock and vulture. Observations such as these have led researchers to argue that our concepts are organized in terms of either best examples or prototypes with a gradient of typicality derived on the basis of similarity to these standards. Rosch (1975) has found that pictures of typical examples (e.g., robins) are identified more rapidly when preceded by a category label (e.g., bird) than in a control condition with no prime (e.g., when robins is not preceded by bird). These positive priming effects do not extend to atypical category members. A straightforward explanation of these results is that activation of the concept *bird* leads to a partial activation or an instantiation of good examples of the category.

More recently, people have become concerned with the importance of context in the structure of categories. Anderson and Shifrin (1980) have taken the extreme view that the meaning of a concept like *bird* is totally dependent upon context. According to this instantiation hypothesis, a category term will be represented in memory as a particular exemplar. Thus, the representation of *bird* in "The bird walked across the barnyard" may be that of a *chicken*.

Although they disconfirmed the strong form of the instantiation hypothesis, Roth and Shoben (1983) found evidence that the typicality of exemplars does vary as a function of context. In their first experiment, subjects saw one of three context sentences: for example, "Mary watched the bird all day," "Mary saw the bird swimming," or "Mary looked at the bird on the telephone wire." Subjects were then timed as they read the test sentence "Mary was very fond of ducks." Roth and Shoben observed faster reading times for the test sentence when the context sentence biased subjects to interpret bird as duck (" . . . the bird swimming") than when the context was relatively neutral (" . . . watched the bird all day"). Moreover, reading time following the neutral context was faster than the reading time following a context that biased subjects against interpreting bird as a duck (". . . on the telephone wire"). Interestingly, exemplars that were judged typical of their category in the absence of explicit context were read more rapidly than atypical exemplars only in the neutral context. These results demonstrate that, at a minimum, category structure must have some degree of flexibility.

A further demonstration of category flexibility comes from a recent study by Barsalou and Sewell (1984). They asked American undergraduates to rate the typicality of category examples for different points of view (e.g., French person, Chinese person, or business person, hippie or housewife). Different typicality gradients for the same categories were obtained from subjects taking different points of view. Although raters agreed with each other for a particular point of view, there was little agreement across points of view for the same category. In other words, people were apparently able to adjust their estimates of typicality or

goodness of example conditional upon the perspective they were asked to adopt.

The ultimate demonstration of the need for concept flexibility, however, may arise when one considers conceptual combinations. One of the basic properties of categories is that they can be combined to form more restricted concepts. Thus, we can restrict birds to songbirds, water birds, predatory birds, and migratory birds. In addition to these combinations, we can construct and understand novel combinations such as gregarious birds or noisy birds. It seems very unlikely that people have a stored representation for each of these conceptual combinations. In addition, adjective–noun pairings such as *friendly–computer* require subjects to create concepts that are more complex than the conjunction of attributes of the two constituents. For example, although most friendly people are warm and outgoing (Wyer & Srull, 1986), it seems silly to describe a computer in this way. Moreover, in conceptual combinations such as *pet fish*, the constituents often have competing values on some dimensions; most pets have fur as their body covering, yet *pet fish* have scales (see Hampton, 1987a, for experiments on property inheritance in conceptual combination). Moreover, it appears, at least on the basis of intuition, that pet fish differ from fish on a number of other dimensions as well as on the dimension of domesticity. For example, pet fish are likely to be both more brightly colored and much smaller than typical fish. Interestingly, our intuitions tell us that these same dimensions are affected differently in pet mice. They are almost certain to be white and to be about the same size as most mice.

Although we will say more later about combinations such as pet fish and pet mice, the present examples illustrate both the flexibility inherent in conceptual combination and the difficulty and complexity in understanding it. Nonetheless, if we want to understand how people structure and use categories, it seems imperative that we deal with conceptual combinations. Because conceptual combination is one of the most basic functions of concepts, it is not surprising that it has received considerable attention recently (e.g., Cohen & Murphy, 1984; Hampton, 1987a, 1987b; Jones, 1982; Oden, 1984; Osherson & Smith, 1982; Smith & Osherson, 1984, 1987; Thagard, 1984; Zadeh, 1982). Probably the most highly developed and explicit model for conceptual combination is the Smith and Osherson (1984) modification model (see also Smith, Osherson, Rips, Albert, & Keane, 1986). Although the model does not claim to provide a complete account of conceptual combination, it has done an excellent job of accounting for typicality judgments involving adjective–noun pairs (see Smith, 1987, for a review). Because this model provides a specific mechanism for dealing with category flexibility (in terms of knowledge restructuring) and because it provides a framework for the experiments we shall be describing, the modification model is presented in some de-

tail. Although the experiments are organized around the modification model, our conclusions apply to a broad class of theories of conceptual combination.

THE MODIFICATION MODEL FOR CONCEPTUAL COMBINATION

The modification model postulates a pair of simple mechanisms by which the meaning of a conjoined category can be derived from the individual meanings of the adjective and noun constituents. The basic idea is that the adjective directs the knowledge restructuring in a straightforward manner by restricting the range of acceptable values and by increasing the importance of the corresponding dimension. If one considers the color of an apple, for example, one notes that most apples are red, but that some are green or yellow and that a few are brown. According to the modification model, for the conjoined category *red apple*, the acceptable values of apple are restricted to red and the dimension of color is given more weight or importance than it has in the simple concept of apple. The typicality of potential instances of the red apple will be a function of their similarity to the newly created prototypic red apple. *Red apple* seems like a fairly familiar concept which may be already directly stored in memory rather than constructed, but Smith and Osherson also deal with less familiar adjective–noun concepts such as *striped apple*. As we shall see, their analysis provides a nice account of some otherwise puzzling results on typicality judgments for conjoined concepts.

To understand just how the modification model works, a more detailed description is necessary. First of all the modification model assumes that the meaning of a noun can be represented as a set of dimensions and a distribution of possible values for each dimension. Smith and Osherson (1984, 1987) assume that these values can be thought of in terms of votes that reflect "intensity and subjective likelihood of appearing in a given instance." For example, if 80% of the apples are red, then 80% of the votes assigned to the color dimension should be given to the value red. The modification model assumes further that dimensions vary in diagnosticity; some are more important than others. Diagnosticity is assumed to vary with the utility of a dimension in making categorizations, but diagnosticity is also allowed to vary dynamically in conceptual combination. The model assumes that when a noun is combined with an adjective, the adjective restricts the range of acceptable values of the relevant dimension (e.g., to red, for red apple) and raises the diagnosticity of that dimension (e.g., color) relative to the situation where there is no adjective.

An example is shown in Table 1. Here, attributes and values are listed for *apple* and *red apple*. Diagnosticities are associated with dimensions and votes are associated with particular values on a dimension. For instance, Table 1 indicates that color has greater diagnostic value than shape and that the subjective frequency of red apples is roughly five

TABLE 1
Distribution of Votes for Apple and Red Apple as Specified by the Modification Model

Apple		Red apple	
Diagnosticity	Dimension	Diagnosticity	Dimension
1	Color	2	Color
	Red 25		Red 30
	Green 5		Green
	Brown		Brown
.5	Shape	.5	Shape
	Round 15		Round 15
	Square		Square
	Cylindrical 5		Cylindrical 5
.25	Texture	.25	Texture
	Smooth 25		Smooth 25
	Rough 5		Rough 5
	Bumpy		Bumpy

Note. Diagnosticities are associated with dimensions and votes with particular values on a dimension.

times greater than the subjective frequency of green apples. There are two differences between the concepts *apple* and *red apple*. First, all of the votes for color have been shifted to red for the red apple concept. Second, the diagnosticity of the color dimension is higher for red apple than it is for apple. Note that the values and the diagnosticity for the other dimensions of shape and texture remain unchanged.

The principal task with which the modification model is concerned is judgments of goodness-of-example. The modification model assumes that this amounts to a similarity judgment, and to derive specific predictions Smith and Osherson (1984, 1987) offer a specific similarity metric. Smith and Osherson assume that similarity is a function of matching and mismatching votes on the various dimensions, weighted by the diagnosticity of the dimensions. This is formally equivalent to Tversky's (1977) well-known contrast rule. For example the similarity between an instance (I) and the features of the apple prototype (A) is given by

$$\text{Sim(I,A)} = f(\text{I\&A}) - f(\text{A-I}) - f(\text{I-A}), \quad (1)$$

where I&A represents the set of votes common to the instance and prototype, A-I designates the set of votes distinct to the prototype, and I-A represents the set of features distinct to the instance. Equally important, f is a function that measures the importance of each set of features and in effect multiplies each vote by the diagnosticity of the corresponding dimension. Similarity is an increasing function of votes common to and a decreasing function of votes distinct to an instance and the corresponding prototype. Goodness-of-example judgments are assumed to be a monotonically increasing function of the overall similarity measure.

We can use our example in Table 1 to illustrate how one derives predictions from the modification model. Let us assume that we are judging the typicality of two pictures, one of a red apple and one of a brown apple. With respect to the concept apple, we can see that most, but not all, votes of the pictured red apple match on the color dimension and presumably on the other relevant dimensions as well. For the picture of the brown apple, very few of the votes on the color dimension match and consequently the typicality of the brown apple is quite low. More extreme results are obtained when we compare these two pictures to the concept *red apple*. Using the newly constructed red apple prototype as the standard to compare with the pictured brown apple, one notes that now all of the votes mismatch on color and, because the diagnosticity of color has been increased, the mismatches count more. Consequently, the typicality of the picture of a brown apple to the concept red apple is very much less than it was for the concept apple. We also obtain somewhat more extreme results for the picture of a red apple. For the color dimension, all of the votes match and these matches count more because of the increase in diagnosticity given to the color dimension. Thus, the modification model has come up with two distinct predictions: a red apple is a slightly better example of a red apple than it is of an apple, and a brown apple is a poor example of an apple, but a much worse example of a red apple. The data are in agreement with these predictions.

Perhaps more impressive is the fact that the model is also able to make correct predictions concerning negatively diagnostic attributes, such as brown. According to classical versions of fuzzy set theory (Zadeh, 1965), the typicality of an object to a combined category ought to be subject to the *min rule*: It ought to be equal to the minimum typicality of the object to its constituents. Thus, for example, the typicality of a picture of a brown apple to the concept brown apple ought to be equal to the minimum of the typicality of the picture of a brown apple to the concepts *apple* and *brown thing*. Therefore, the fuzzy set theory implies that the typicality of a brown apple as a *brown apple* cannot exceed its typicality as an apple.[1] The modification model, however, does not make this prediction. The picture of a brown apple will have few mismatching votes with the concept brown apple and consequently that typicality will be quite high. In contrast, the numerous mismatches on the color dimension will ensure that the typicality of our pictured brown apple to the concept

[1] This prediction of fuzzy set theory hinges on the assumption that adjective–noun pairs are interpreted by conjoining distinct categories and that typicality judgments for combined concepts are on the same scale as typicality judgments for sample concepts. Both assumptions have been questioned (e.g., Oden, 1984; Jones, 1982). These issues are, however, orthogonal to our present concerns.

apple will be very low indeed. Thus the modification model predicts violations of the min rule.

Data examined by Smith and Osherson (1984, 1987) confirm this prediction of the modification model. They found numerous violations of the min rule and also noted that the typicality of an instance in a conjoined concept often exceeded the maximum of its typicality in the constituent concepts. Thus, it would appear that the modification model provides a relatively straightforward mechanism that accounts for numerous findings on category flexibility and conceptual combination.

LIMITATIONS OF THE MODEL

Although models such as the modification model are appealing for their simplicity, we believe that they are seriously limited as models of conceptual combination in particular and knowledge restructuring in general. In our view, the main problem is the assumption that the dimensions corresponding to adjectives in adjective–noun combinations are independent of one another. An alternative idea is that concepts have an internal structure based on a variety of interproperty relationships and that, as a consequence, constituent dimensions are not independent of each other (e.g., Medin, Wattenmaker, & Hampson, 1987; Murphy & Medin, 1985; Wattenmaker, Dewey, Murphy, & Medin, 1986). According to this view, when a value on some dimension is specified, the value on other dimensions also may be changed. For example, our intuition is that a brown apple is less likely to be shiny and more likely to contain a worm than is a red apple. Other studies show that people are sensitive to within-category correlated attributes (e.g., Malt & Smith, 1984; Medin, Altom, Edelson, & Freko, 1982) and we think this sensitivity will extend to typicality judgments of adjective–noun combinations. Borrowing an example from Cohen and Murphy (1984), a bald, old man may be judged to be a more typical man than a bald, young man even though an old man may be no more typical of man than is a young man. Our first study evaluates the role of correlated attributes in typicality judgments.

We believe that interproperty relationships are structured such that the meaning of individual adjectives might change across noun contexts. For example the adjective *gold* might have its greatest impact on the dimension of value in going from *coin* to *gold coin* but might have a greater effect on the dimension of color in going from *railing* to *gold railing*. Similarly we believe that the understanding of adjectives like *baked, boiled,* and *fried* varies depending on whether one is talking about potatoes or fish. The second study is concerned with this possibility.

Finally, if concepts have an internal structure then properties of concepts may differ in their centrality. That is, the same property may be equally true of two different concepts but may be more central to one of

the concepts than to the other by virtue of the role it plays in the internal structure of the concept. For example, basketball and cantaloupes are both round but roundness may be a more central aspect of basketballs than cantaloupes. The third study examines property centrality.

Although it is convenient to organize our experiments in terms of limitations of the modification model, our intention is not to single out the modification model for criticism. The modification model assumes that category flexibility is achieved through a straightforward form of knowledge restructuring. It is easier to organize our studies in terms of needed extensions of the modification model than to try to lay out a complete account of category flexibility. We claim that theories of conceptual knowledge organization will, in general, require much more restructuring than almost all current theories imply.

EXPERIMENT 1

The modification model assumes that the noun representation in adjective–noun pairs is modified only with respect to the dimension associated with the adjective. As a consequence, the model is insensitive to any effects of correlated attributes. For example, it postulates that the representation of wooden spoon is identical to the representation of spoon with the exception of the votes on the materials dimension. In contrast, we suspect that people will take the fact that a spoon is wooden as implying something about the values on other dimensions such as size and function.

If the adjective of an adjective–noun pair modifies not only value and diagnosticity on the selected dimension but also modifies correlated dimensions, then these modifications should be reflected in typicality judgments for combined dimensions. For example, let us consider the typicality of exemplars such as wooden spoon and metal spoon as exemplars of both the simple category *spoon* and the combined category *large spoon*. Because the modification model assumes that combining concepts affects only a single dimension, both wooden spoon and metal spoon differ from *spoon* only in their votes on the materials dimension. Similarly, large spoon differs from *spoon* only on the size dimension. Consequently, if we find that metal spoon is a better example of *spoon* than is wooden spoon, then it must be the case that metal spoon is also a better example of *large spoon*, because metal spoon and wooden spoon both mismatch to the same degree on the dimension of size, but wooden spoon must have more mismatches on the material dimension because it is a poorer example of *spoon*. In addition to this independence prediction, the modification model is constrained to make a number of other interesting predictions that we will consider shortly.

Method

Subjects were asked to judge the typicality of combined concepts with respect to a number of categories. For each noun, two pairs of adjectives were selected where each member of the pair denoted one extreme of one dimension. For example, the pairs busy–empty and paved–unpaved were used with street. We attempted to select dimensions with correlated values (e.g., busy streets should be more likely to be paved than nonbusy [empty] streets).

Materials. Twenty nouns were selected and each was paired with two sets of two adjectives. All possible combined concepts were formed and subjects judged the typicality of each combined category with respect to three categories: the simple noun category and each of the combined categories formed by pairing each adjective from the other dimension with the noun. For example, subjects judged the typicality of wooden spoon with respect to *spoon, large spoon,* and *small spoon.*

Procedure and design. For each noun, there were four category combinations and three typicality ratings, making 12 possible judgments. Subjects made only half of these 12 typicality judgments; one group of subjects judged only the typicality of the combined categories derived from the two adjectives from one dimension (e.g., metal and wooden spoons as spoons, small spoons, and large spoons) and a second group made the same judgments for the combined categories derived from the other dimension (e.g., small and large spoons as spoons, metal spoons, and wooden spoons).

Subjects were given a booklet in which to make their ratings. The ratings were blocked such that subjects made all six typicality judgments with respect to one noun (and its combined derivatives) at one time. These nouns were randomly assigned to pages and the ordering of these pages was randomized for each subject.

Subjects. The subjects were 26 undergraduates at the University of Illinois who participated as part of a course requirement. Thirteen were assigned to each of the two groups. One subject failed to follow instructions so data were obtained for only 25 of the participants.

Results

The mean typicality ratings for the various conceptual combinations are shown in the Appendix. Recall that a given subject contributed ratings to only 6 of the 12 possible combinations associated with a noun. Overall, there were few asymmetries in the judgments. For example, a small spoon as a *metal spoon* had a mean rating of 7.75 compared with a rating of 7.14 for a metal spoon as a *small spoon.* The main results on correlated attributes were clear-cut. For example, for the target concept *small spoon,* metal spoons were rated as far more typical than wooden spoons (7.75 versus 2.33), but this pattern reversed when the target concept was *large spoon* (7.25 versus 4.67). The averaged results for this example are presented in Table 2. Because the use of mean ratings makes inappropriately strong assumptions about the underlying scale (that it is at least an interval scale), our analysis will focus on ordinal precautions of the modification model and will treat the judgments as ordinal.

Before examining these ordinal constraints, however, we note that the mean typicality ratings violate the relationships predicted by the modifi-

cation model. Of particular interest is the independence rule outlined earlier. Here the results are clear-cut: For 19 of our 20 noun categories, we find violations of the independence rule (e.g., if metal spoon is more typical of *spoon* than is wooden spoon, then it should be more typical of *small spoon, large spoon, green spoon,* and so on). Our results are, of course, highly reliable by a sign test ($p < .01$).

The explanation for this finding is that we were successful in our attempt to find correlated dimensions. In the spoon example, wooden spoons are used for stirring food, particularly in circumstances where one is worried about scratching the cooking vessel. On the other hand, metal spoons are used not only for stirring, but also for eating. Consequently, while one may have both large and small metal spoons, one is likely to have only large wooden spoons. In other words, size and material are not orthogonal, and consequently we do not observe independence in these judgments of typicality.

In addition to this prediction, the modification model makes a number of other ordinal predictions that can be tested against the present data. In many respects, we can think of these predictions as an extension and formalization of the prediction we just examined. Let us describe them in terms of our spoon example. We can state constraints associated with the modification model as equations. First,

$$\text{ws(S)} < \text{ms(S)} \Rightarrow \text{ws(SS)} < \text{ms(SS)}, \tag{2}$$

where ws(S) is the typicality of a wooden spoon of a spoon and ms(SS) is the typicality of a metal spoon as a small spoon. That is, if the typicality of a wooden spoon as a spoon is less than the typicality of a metal spoon as a spoon, then the typicality of a wooden spoon as a small spoon should be less than the typicality of a metal spoon as a small spoon. This prediction of the modification model comes from the fact that the size votes for wooden spoon and metal spoon are assumed to be identical. For similar reasons,

$$\text{ws(S)} < \text{ms(S)} \Rightarrow \text{ws(LS)} < \text{ms(LS)}, \tag{3}$$

where ms(LS) is the typicality of a metal spoon to a large spoon. Finally, similar reasoning leads to the prediction

$$\text{ws(SS)} < \text{ms(SS)} \Rightarrow \text{ws(LS)} < \text{ms(LS)}. \tag{4}$$

TABLE 2
Mean Typicality Ratings for Conceptual Combinations of *Spoon*

	Wooden	Metal
Large	7.74	4.68
Small	2.44	7.44

Notice that these three predictions are not independent; if (2) and (4) are true then (3) must be true by transitivity.

In addition to these three predictions, there are also three other predictions that can be made. Like the first three, Eqs. (5) through (7) also stem from the modification model's assumption that attributes are independent,

$$\text{ws(SS)} < \text{ws(LS)} \Rightarrow \text{ms(SS)} < \text{ms(LS)} \quad (5)$$

$$\text{ws(S)} < \text{ws(SS)} \Rightarrow \text{ms(S)} < \text{ms(SS)} \quad (6)$$

$$\text{ms(S)} < \text{ms(LS)} \Rightarrow \text{ws(S)} < \text{ws(LS)}. \quad (7)$$

Let us describe the rationale for Eq. (5) in some detail. Recall that spoon has some distribution of values on the dimensions of material and size. According to the modification model, changing from spoon to wooden spoon alters only the votes on the materials dimension. Consequently, if ws(SS) < ws(LS), it must be the case that the size dimension of spoon (and also of wooden spoon and metal spoon) has more votes on large than on small. Thus, because the size votes are identical on wooden spoon and metal spoon, it must be true that ms(SS) < ms(LS).

Given these six predictions of the modification model, we can examine our data to see how many violations occur. Before doing so, however, we should consider how to deal with random measurement error. It would not be prudent to reject a model if we found some small number of violations that might have arisen because of experimental error.

In order to adopt a more reasonable criterion, we first determined what a chance number of violations would be. This problem is not trivial because the first three constraints are not independent of each other. However, it turns out that the expected proportion of violations over all six predictions is 50%.[2] Thus, it is very reasonable to claim that, according to the modification model, the number of violations observed in the data

[2] The chance probability for Eqs. (4) through (6) is .5 so we expect 1.5 violations by chance. For the first three equations, let us consider the probability of all possible outcomes. For example, the probability of (1 & 2 & 3) is .25, because the probability that Eq. (1) is true is .5 and the probability that Eq. (3) is true is .5, and if both these equations are satisfied then the probability that Eq. (2) is satisfied is certainty. We thus have a .25 probability of having zero violations. The probability of having one violation turns out to be zero because of this same interdependence. If any two are satisfied, then the third must also be true. However, it is possible for any single one (only) to be true. Thus, for example, if Eq. (1) is true and Eq. (2) is false, then Eq. (3) must be false or it will violate the rule given above. Thus the probability is again .25 as the probability that Eq. (3) is false is 1.0. The same logic holds for all three possibilities of obtaining two violations. Consequently, because the probability of two violations sums to .75 and the probability of zero violations is .25, the expected value is 1.5 violations out of the three interdependent predictions.

should be between zero and what chance would predict, and presumably closer to 0 than 50%.

We examined this prediction by computing the number of violations and agreements associated with Eqs. (2)–(7) for each noun in our list. We also performed the same analysis for each subject. Ties were excluded. On average, we found 63% violations and 37% agreements—more violations than even chance would expect. This tendency for greater than 50% violations was remarkably consistent. Eighteen of our 20 target nouns showed the effect with one having exactly 50% violations and one 39%. The analysis over subjects yielded an identical pattern: Twenty-three subjects had more than 50% violations, one had exactly 50% and one had 39%. Both the results over items and the results over subjects were highly reliable by a sign test ($p < .01$).

These findings contrast strikingly with the predictions of the modification model. Our data clearly violate the constraints expected by the model and suggest that any adequate explanation of combined categories must make provisions for correlated attributes.

Discussion

The results show that specifying the value on one adjectival dimension leads to changed typicality judgments for adjective noun pairs involving a correlated adjectival dimension. Thus subjects judge metal spoons to be more typical spoons than wooden spoons but judge metal spoons to be less typical than wooden spoons of the concept large spoons. In other words, typicality judgments reflect sensitivity to correlated attributes.

We found little evidence for asymmetries as a function of adjective order. Thus, a small spoon as a metal spoon received roughly the same rating as a metal spoon as a small spoon. The largest asymmetry was that brown grass as short grass was rated to be much more typical than short grass as brown grass (6.61 vs 3.83). It is possible that this reflects causal reasoning which is directional; whatever causes grass to be brown also probably causes it to be short, but the converse is not true. We will say little more about asymmetries for now because they were rare and are not of primary concern.

At a minimum, our results suggest that the modification of knowledge associated with changes in one attribute leads to changes in correlated attributes. How serious this finding is for models that treat dimensions as independent depends on the ubiquity of correlated attributes. If correlated attributes are a rare occurrence, then they might appropriately be treated as a special case. For example, one might assume that the representation includes specific stored linkages marking two dimensions as "correlated."

We do not think that correlated attributes can be treated as an excep-

tional case. First of all, it seems to us that correlated attributes are quite widespread. Second, although in some cases attribute correlations may be prestored, it is likely that most of them are computed or generated in the context of comprehension. For example, in thinking about the size of birds, one can readily imagine that size is correlated with color, whether or not a bird sings, type of foot, whether or not the bird migrates, and even the color of eggs. It would impose a seemingly overwhelming burden to assume that all such correlations are prestored. A more likely possibility is that such correlations are computed by retrieving and analyzing examples in the spirit of the Kahneman and Miller (1986) norm theory. We consider this idea in greater detail under General Discussion. For the moment, we turn to a second implication of the claim that an adjective in an adjective–noun pair leads only to a vote shift and a change in diagnosticity on the associated adjectival dimension.

EXPERIMENT 2

Whereas Experiment 1 demonstrated an effect of varying the adjective on the rated acceptability of the adjective–noun combination, Experiment 2 examined the effects of combining a particular triad of adjectives on different nouns. In particular, we asked subjects to judge the similarity of combined concepts such as *brass railing, gold railing,* and *silver railing* and compared these ratings to the ones obtained for *brass coin, gold coin,* and *silver coin.* For this example our intuition is that for railings, brass and gold will be judged as the most similar pair but that for coins, silver and gold will be the most similar pair.

The modification model does not predict that the pattern of similarity judgments will change across noun contexts. In the most straightforward interpretation of the modification model, brass, silver, and gold will all produce mismatching votes on the materials dimension and any similarity will derive from matches on other dimensions. Consequently, all three pairs will be equally similar. One might object that this assumption is too strong in that it implies that white is no more similar to gray than it is to black. This criterion could be relaxed in two ways. First, one might assume that there is some decomposition of values into more primitive dimensions. This assumption would allow for similarities among colors. Alternatively, one might assume that associated with each value is a distribution of votes, centered at the value named but overlapping with similar other values. For example, *white* might be represented by mostly white votes, a few gray votes, and no black votes. Under this assumption white and gray will be more similar than white and black. Although both the decomposition assumption and the distributional assumption have the virtue of allowing some pairs to be more similar than others, they still do not predict an effect of noun context. White should be more similar to

gray than to black regardless of whether one is describing clouds, hair, or bears. By the same token, the similarity relationships found for brass, gold, and silver railings ought to be the same ones observed for brass, gold, and silver coins.

Method

Subjects were given three pairs of conjoined categories and asked to determine which was most similar. For example, subjects were given the pairs brass railing–silver railing, silver railing–gold railing, and brass railing–gold railing and were asked to determine which was the most similar pair and which was the least similar pair.

Materials and design. There were 16 triplets of adjectives that were selected such that they could all plausibly be applied to at least several nouns. Each triplet of adjectives was judged in the context of three different nouns. A complete list of the materials is presented in Table 3.

Subjects rated each triplet of adjectives in only one noun context. Thus noun context was a between-subject factor and adjective triplet was a within-subject factor.

Procedure. Subjects were instructed to judge the similarity of three pairs of conjoined concepts and to indicate which pair of concepts was the most similar and which was the least similar. All subjects rated 16 such triads which were presented in randomized order. The experiment lasted about 15 min.

Subjects. The subjects were 46 University of Illinois undergraduates who participated as part of a course requirement.

Results

The principal theoretical question of interest is whether the relations among the adjectives remain constant across contexts. If the meaning of a conjoined concept can be characterized as some combination of the meaning of its two constituents, as predicted by the modification model, then we ought to find no effect of context.

In striking contrast to this prediction, we found that the relations among the adjectives changed rather dramatically as a function of the associated noun. To take our earlier example first, although brass railing and gold railing were judged the most similar in the context of railings, gold coin and silver coin were judged the most similar in the context of coins. In other words, brass and gold were more similar than either brass and silver or gold and silver in the context of railings, but silver and gold were more similar than brass and gold or brass and silver in the context of coins. More generally, we found this kind of difference for a large majority of our items. The results are summarized in Table 3, which shows the proportion of time an adjective pair was selected as the most similar for each noun context based on mean ranks. The 16 triplets of adjectives were treated as independent tests. The changes as a function of noun contexts were highly reliable by a χ^2 test ($\chi^2 = 106.77$, $df = 64$, $p < .001$). Only 3 of the 16 triplets failed to produce a shift of at least 20 percentage points, and 4 triplets had shifts of 40 percentage points or

TABLE 3
Proportion of Time a Given Adjective Pair Was Selected as Most Similar as a Function of Noun Context for Each of the 16 Triplets of Adjectives

Target concept	Adjective pair		
	Brass, gold	Brass, silver	Gold, silver
Railing	.50	.07	.43
Ring	.41	.00	.59
Coin	.27	.07	.66
	Copper, iron	Copper, steel	Iron, steel
Pot	.14	.00	.86
Pipe	.18	.06	.76
Bowl	.20	.07	.73
	Black, blue	Black, green	Blue, green
Birds	.71	.00	.29
Shoes	.88	.00	.12
Eyes	.13	.00	.87
	Checkered, spotted	Checkered, striped	Spotted, striped
Flag	.86	.14	.00
Animal	.59	.23	.18
Shirt	.40	.60	.00
	Metal, plastic	Metal, wooden	Plastic, wooden
Spoon	.36	.28	.36
Crate	.47	.18	.35
Desk	.27	.40	.33
	Brown, red	Brown, yellow	Red, yellow
Leaves	.50	.21	.29
Apple	.29	.12	.59
Sunset	.33	.00	.67
	Baked, boiled	Baked, fried	Boiled, fried
Eggs	.14	.50	.36
Potatoes	.47	.29	.24
Fish	.20	.33	.47
	White, gray	White, black	Gray, black
Cloud	.43	.00	.57
Bear	.23	.00	.77
Hair	.73	.00	.27
	Full-time, part-time	Full-time, temporary	Part-time, temporary
Hobby	.14	.00	.86
Coach	.40	.13	.47
Job	.18	.06	.76
	Canned, fresh	Canned, frozen	Fresh, frozen
Peas	.07	.50	.43
Fish	.06	.35	.59
Fruit	.07	.73	.20
	Drama, history	Drama, science	History, science
Book	.50	.00	.50
Major	.35	.00	.65
Building	.33	.00	.67

TABLE 3—*Continued*

Target concept	Adjective pair		
	International, national	International, local	National, local
Weather	.50	.00	.50
News	.65	.00	.35
Crime	.67	.00	.33
	Cotton, silk	Cotton, velvet	Silk, velvet
Blanket	.00	.21	.79
Nightgown	.12	.17	.71
Blouse	.20	.13	.67
	Brick, concrete	Brick, gravel	Concrete, gravel
Road	.64	.29	.07
Barricade	.94	.00	.06
Wall	.67	.06	.27
	Ceramic, glass	Ceramic, plastic	Glass, plastic
Tile	.64	.29	.07
Bottle	.71	.00	.29
Mug	.73	.27	.00
	Cotton, nylon	Cotton, woolen	Nylon, woolen
Socks	.21	.72	.07
Underwear	.47	.35	.18
Gloves	.13	.87	.00

more. Clearly, the similarity judgments were not independent of noun context.

Discussion

Our results show major effects of noun context on the similarity judgments associated with adjective–noun pairs. The modification model, in its present form, is unable to predict these interactions. There is, however, an interpretation of the modification model that can account for some but not all of the results. We first describe this interpretation and then argue that it proves to be inadequate.

Simple versus complex adjectives. One could argue that some of our adjectives are not associated with a single underlying dimension but rather multiple component dimensions.[3] For example, the materials dimension associated with brass, gold, and silver can be broken down into constituents of color, value, malleability, and a variety of other dimensions. If we also assume that these dimensions have diagnosticities that vary from concept to concept, then we can account for the railing versus

[3] We are indeed ignoring distinctions among adjectives that for other reasons may be quite important, (Levi, 1978). We believe that the problems associated with models that attempt to describe combined categories by adding or changing a single feature are quite general.

coin results. That is, brass and gold are most similar in the context of railings because the dimension of color has high diagnosticity relative to the dimension of value. This relationship among diagnosticities is reversed in the context of coins, and consequently gold and silver are now the most similar pair.

Although something like the above account may work for gold, silver, and brass railings versus gold, silver, and brass coins, it seems too limited to account for our full set of results. We think that at least three other factors are entering into the similarity judgments: (1) correlated attributes, (2) similarity of typicality, and (3) causal relationships.

Correlated attributes. The first study showed that typicality judgments are influenced by correlated attributes. If there are correlated attributes associated with adjective–noun pairs, then they ought to change similarity relations. For example, in judging the similarity of metal, plastic, and wooden spoons, the fact that wooden and plastic spoons tend to be large should add a dimension of similarity between wooden and plastic spoons and a dimension of difference between these two types of spoons and metal spoons. This correlation should (and did) lead plastic and wooden spoons to be judged to be the most similar adjective–noun pair for *spoon* but not for *crate* or *desk*.

Similarity of typicality. If people instantiate adjective–noun pairs and consider whether or not there are real-world examples of them, then one might expect that judgments will be influenced by a form of second-order similarity, that is, by similarity with respect to typicality. For example, penguins and peacocks are not especially similar but they do share the fact that they are rather atypical or odd birds. For the triple cotton, nylon, and woolen underwear, people may judge cotton and nylon underwear to be the most similar pair because they know that cotton underwear and nylon underwear are more common than woolen underwear. One might be able to interpret this similarity of typicality factor as a special case of the correlated attributes principle where the correlated property is typicality or frequency of occurrence.

Causal relations. In our view a serious limitation of the modification model is that it does not consider the role of theoretical knowledge and causal relations in structuring concepts and in interpreting adjective–noun combinations. Consider, for example, white, gray, and black clouds versus white, gray, and black hair. In the case of clouds, one might consider white clouds to be the normal situation with gray and black clouds associated with a change of state to stormy conditions. Whatever the cause of storm conditions is, black and gray clouds are associated with it in a way that white clouds are not. Therefore, one might expect gray and black clouds to be judged to be more similar than white and gray clouds (and that proved to be the result). The situation is reversed when the

noun concept is hair. For hair, black may be viewed as the normal condition with gray and white representing changes of state associated with the aging process. Therefore, white hair and gray hair ought to be (and were) judged to be more similar than gray and black hair. In other words, people may not be rating simply the similarity of colors in context but also the similarity of their function (e.g., as cause or effect) within a phenomenon defined by the context.

If causal and theoretical relations influence judged similarity, as we have argued, then the type of representation associated with the modification model is not sufficiently rich to describe the manner in which people interpret combined concepts. The third study examines the idea that a property may be equally true of two different concepts, but by virtue of underlying causal relations, may be more important or central in one concept than in the other.

EXPERIMENT 3

The goal of Experiment 3 was to determine if combined concepts varied in their importance or centrality, and thus in their typicality, even when they did not vary in frequency. Our basic logic or strategy was as follows: We generated concept properties that we thought would always or nearly always be true of two different concepts (and thus equal in subjective frequency or votes) but perhaps more important in one concept than another. For example, our intuition is that almost all bananas are curved and almost all boomerangs are curved. (Actually, experts tell us that there are apparently three types of boomerangs, only one of which is curved—but that is irrelevant to our arguments.) Although one might be able to come up with a reason for the curvature of bananas, the rationale for the curvature of boomerangs is immediately obvious. To produce items for typicality judgments, we then negated the property (curved, in our example) for each of the concepts. Thus, subjects were asked to judge the typicality of a straight banana as a banana and the typicality of a straight boomerang as a boomerang. If a property is more central to one concept than another, then one ought to see differences in typicality judgments.

In other words, we sought to find combined concepts that had a frequency of near zero, such as *straight banana, square cantaloupe, flying ostrich,* and *square basketball,* but which varied in terms of the degree to which they were still members of the noun category. For example, we felt that the shape of a basketball was central to the meaning of the concept. If a basketball is not round, then it cannot be a basketball because it cannot be dribbled, passed, or shot in the manner that is characteristic of the game. On the other hand, we believed that shape was relatively less important for a cantaloupe. Although square cantaloupes never occur in

nature, by our intuitions there is nothing essential about the shape of a cantaloupe. It must have flesh, seeds, and a rind, but even a peculiarly shaped cantaloupe is still a cantaloupe.

Without additional assumptions, the modification model has no basis by which to account for effects of centrality. The modification model allows for differences in diagnosticity or weighting of dimensions, but diagnosticity presumably is based on how useful or informative a dimension is in classification. In this sense diagnosticity corresponds to cue validity or predictive value. We shall argue that centrality cannot be equated with diagnosticity.

Method

We first gave one group of subjects a booklet of adjective–noun pairs and asked them to judge their frequency. We gave the same list to a second group of subjects and instructed them to rate the adjective noun pairs on an 11-point scale in terms of the degree to which a noun that possessed the quality indicated by the adjective would still be a member of the noun category.

Materials. The stimuli were 30 adjective–noun pairs in which a single adjective was paired with two nouns. For each, the authors agreed that the adjective was more central to one noun than to the other (as in square basketball and square cantaloupe).

Procedure. One group was instructed to estimate the number of times (out of 100) that the noun possessed the attribute indicated by the adjective. Subjects in this group were given examples and instructed to be as accurate as possible. The second group was instructed to rate the degree to which the adjective–noun pair belonged to the category indicated by the noun. They made this rating on a 10-point scale and examples were again provided. For both groups the order of items was randomized. Each task required about 15 minutes to complete.

Subjects. The subjects in this experiment were students in classes at the University of Illinois who volunteered to participate in the study. Altogether, 24 subjects participated in the frequency group and 14 in the degree-of-belonging group.

Results

As our goal was to examine rated differences in category membership when the frequency was zero, we adopted a strict criterion for including an item in our analysis. More specifically, we required that at least 20 of our 24 subjects rate the frequency of both members of the pair as 0 or 1 out of 100. To demonstrate that this criterion is strict, many of our pairs which never occur (in our own experience) failed to meet this criterion: for example, plastic fence posts, windowless mansions, black bridal gowns, and opaque Coke bottles.

The 15 pairs that did meet our criterion are listed in Table 4 along with their mean rating of the degree to which they belong to the category. For each pair, the noun for which we felt the adjective was most central is the second adjective–noun pair listed. Although we have reported the mean in Table 4 because it is the most widely understood descriptive statistic, we were reluctant to assume interval measurement in the ratings and

TABLE 4
Number of Subjects out of 24 Judging that a Given Pair Had a Real World Frequency of 0 or 1 out of 100 along with Goodness-of-Example Ratings for Experment 3

Pair	Goodness rating	Frequency judgment
Pink money	3.71	20
Pink grass	2.57	22
Polka-dot fire hydrant	7.43	21
Polka-dot yield sign	3.71	23
Soft knife	2.79	23
Soft diamond	1.42	22
Square cantaloupe	2.36	24
Square basketball	0.21	24
Straight banana	3.79	23
Straight boomerang	1.14	23
Striped lemon	2.57	22
Striped sun	2.14	22
Triangular capsule	3.29	21
Triangular record	2.36	22
Two-wheeled car	2.43	22
Two-wheeled truck	2.43	22
Waterless melon	2.86	21
Waterless ocean	0.79	23
Wooden doormat	4.93	21
Wooden skillet	1.93	24
Yellow cherry	4.29	21
Yellow blood	1.86	22
Diamond-shaped pie	6.43	20
Diamond-shaped stop sign	3.21	23
Flying ostrich	2.00	23
Flying whale	1.21	24
Green swan	3.21	24
Green clouds	2.29	22
Orange Girl Scout uniform	5.36	23
Orange coffee	3.43	21

Note. The ratings were on an 11-point scale (0 thrugh 10) with higher numbers corresponding to better examples.

hence examined the number of subjects who rated the bottom adjective–noun combination higher than the top adjective–noun combination. For 14 of the 15 pairs listed in Table 4, more subjects judged the top combination as a better example of the noun category ($p < .01$, by a sign test).

Thus, there are adjective–noun pairs that never occur (as defined by our criterion) yet that vary in terms of the degree to which they belong to the category. Contrary to the modification model which assumes that the distribution of votes is based on subjective frequency and that diagnosticity is based on cue validity, we find that there are numerous examples of adjectives that are never true of the paired noun, but which nevertheless vary in terms of the degree to which they are acceptable examples of the noun category.

Discussion

We found clear evidence for property centrality effects that cannot be explained in terms of property frequency. One way in which the modification model might account for the results of this experiment is to claim that *straight banana* is a better example of *banana* than *straight boomerang* is of *boomerang* because there are more other dimensions with matching values for bananas and straight bananas than there are for boomerangs and straight boomerangs. Thus, although there might be mismatching votes on the shape dimension for both noun concepts, there are more matching ones for banana than for boomerang. Although this formulation will work mechanically, it makes the counter-intuitive prediction that the more one knows about a concept the more willing one is to accept an exception. Furthermore, our data do not appear to line up in a pattern based on likely amount of knowledge. For example, we think that most people know at least as much about whales as they know about ostriches.

An alternative way to salvage the modification model is to argue that effects that we attribute to centrality can be explained in terms of diagnosticity. Although this explanation may seem plausible in the case of curvature in boomerangs versus bananas, it seems less plausible for a dimension such as color in *yellow cherry* and *yellow blood* or for a dimension such as mode of locomotion in *flying ostrich* and *flying whale*. Although we do not have direct empirical evidence for our claim, it would seem to us that mode of locomotion would be at least as diagnostic for whale as for ostrich.

Related to the idea of varying diagnosticities is the idea that the number of votes associated with a particular value may vary as a function of the concept. For example, one might claim that curved received more votes in the context of boomerangs than it does in the context of bananas. This explanation of our results depends on the empirical measure one uses to assess the number of votes. Smith et al. (1986) used attribute listings and with this procedure it might prove that our central attributes of Experiment 3 would receive more listings than our less-central ones. However, we would argue that attribute listings are themselves likely to be influenced by centrality and diagnosticity and thus differences in the

frequency of listings would reflect centrality, at least indirectly. Our own preference for assessing the number of votes is to ask people to estimate frequency explicitly, which seems a more direct way of assessing the number of votes for each value.

From these data, it appears that more radical approaches to conceptual combination are needed. For example, one might assume that there is structure among attributes that guides the reasoning about adjective–noun pairs. Consider again flying whales versus flying ostriches. For *flying whale*, one must make changes to a large number of attributes. One must add some kind of wings, slim down the body considerably, and find some means of propulsion to get our redesigned whale into the air. In contrast, a *flying ostrich* requires fewer and less drastic changes. Shortening the legs and neck, for example, might very well enable this hypothetical ostrich to fly. People who rated green swans as typical swans often mentioned the idea of selective breeding and argued that one could create green swans that would be perfectly acceptable swans. This line of thinking is most compatible with the Kahneman and Miller (1986) norm theory that relies on the notion of mutability and possible-worlds reasoning. We return to these ideas under General Discussion.

GENERAL DISCUSSION

The results from our three experiments indicate that models that attempt to explain combined categories by adding or changing a single feature are not successful. Attributes are not independent of each other, and people are sensitive to correlations among attributes. Even something as simple as an adjective similarity judgment interacts with the particular noun in an adjective–noun pair, and the same property that may be equally true of two concepts is more central to one concept than the other. It appears instead that any formulation must rely heavily on the relationships among attributes and on the multidimensional effects that adding a qualifying attribute has on a noun.

This generalization is true even for a common and seemingly uncomplicated adjective like *large*. Let us consider the concepts *tropical fish* and *large tropical fish*. Following the modification model, one would think that the change from the simple concept to the combined one would involve a change solely on the dimension of size. However, to a person who has kept tropical fish it seems obvious that large tropical fish differ from tropical fish in a number of ways. Large tropical fish are generally more pugnacious; they are less suitable fish to keep in a community tank. They are more likely to have a coloring pattern that will change with age. They are more likely to be monogamous and to be caring parents. They are much more likely to be oviparous than tropical fish. Large tropical fish exhibit more territoriality and engage in more digging behavior. They

also have a different diet and have different water requirements in terms of pH and degrees hardness.

It should be noted that all of these attributes are not correlated with largeness in the same way. Although it it generally true that size and ferocity are correlated in the real world, the relation between size and digging behavior appears to be unique to tropical fish. The importance of this distinction is that one cannot solve the problem of correlated attributes simply by postulating a network of interconnections among attributes independent of associated nouns.

Of course, one might wonder whether the attribute *large* is representative of all adjectives. We attempted to come up with some attributes that did not affect values on other dimensions. Our best candidates are adjectives of color: the examples used by Smith and Osherson (1984). Even here, however, it appears that there may be some effects on other dimensions. Thus, for example, a red bird is less likely than a bird to be large, a red apple is less likely than an apple to be wormy, and a red tennis ball is less likely than a tennis ball to be cheap.

Furthermore, even what appears to be simple adjectives such as color terms are embedded in causal relations that alter similarity judgments. Thus, white hair and gray hair are more similar than gray hair and black hair, but white clouds and gray clouds are less similar than gray clouds and black clouds. In the noun context of hair, white and gray are linked by our understanding of the aging process, whereas in the noun context of cloud, gray and black are linked by our understanding of storm conditions. Again, the upshot of this finding is that adjective–noun conjunctive concepts do not involve solely a selection on the associated adjective dimension. A final blow to the idea of treating component dimensions of concepts in an independent manner is the observation that properties differ in their centrality where centrality cannot be equated with either the probability of a property being present or its variability. For example, neither whales nor ostriches fly, but flying ostriches are more typical of the concept ostrich than flying whales are of the concept whale.

Theoretical Implications

Models that assume a prestored representation for noun concepts have a great deal of difficulty accommodating the changes than can occur as a result of a combination with a particular adjective. The main problem is that too much information must be prestored. In view of the observation that many if not most attribute pairs are not orthogonal, one would need to store an excessively long list of attribute correlations. One might attempt to produce some economy by assuming that some relatively context-free inference rules about potential correlations are stored. Many such inferences, however, would be context-dependent. For example, in-

creased size is often associated with increased longevity, but this relationship is reversed for dogs. Thus a model that assumes prestorage of base information cannot handle this phenomenon simply with a general rule that changes on the size dimension change the values on the longevity dimension in a predictable and invariant way; instead, any adequate model must be able to use specific world knowledge to obtain the correct relationships. That is, information concerning attribute correlations frequently must be computed rather than prestored and the form of concept representation must be able to support such computations.

Given these constraints, it is hard to imagine a semantic network that could capture the complexity and the context dependence of adjective–noun relationships. Such a structure must have a mechanism for determining both the centrality of a particular adjective in varying contexts and also the relationships among dimensions in varying contexts. We see no obvious way to formulate prototype or feature-based models of abstraction that can accomplish these tasks.

Perhaps the problem with prototype or feature-based models of abstraction is that they assume too much abstraction too early. In Estes' (1986) terms, perhaps filtering of examples is assumed to occur too early in the abstraction process. Perhaps more specific exemplars of base concepts are accessed in memory, and consideration of these exemplars allows us to generate the combined categories. We will return to this possibility in a moment.

Types of knowledge reorganization. By now it should be clear that our results pose problems not only for the Osherson and Smith modification model, but also for a large class of models for knowledge restructuring. We first briefly review five types of models for knowledge restructuring and then evaluate then in light of our results.

1. *The refocusing hypothesis.* One interpretation of the original Roth and Shoben (1983) findings concerning context effects on typicality is that context determines what the best example will be and that typicality depends on similarity to this best example. To test this idea, Roth and Shoben presented subjects with pairs of sentences in which the best exemplar did not change. For example, in the sentence "The secretary enjoyed her beverage every morning during her break," the best example of beverage is *coffee*. Similarly, for the sentence "The truck driver enjoyed the beverage with his doughnut" *coffee* is again the best exemplar. If, as predicted by the refocusing hypothesis, the relationship among beverage exemplars does not change as a function of context, then the ordering of acceptability of exemplars in these two scenarios should be identical. Given two other exemplars, *tea* and *milk*, for example, if *tea* and *coffee* are more similar to each other than *milk* and *coffee*, whenever coffee is the best example, tea will also be a better example of beverage than milk.

This prediction was disconfirmed by Roth and Shoben in two experiments. In the first, subjects simply rated the acceptability of completions such as "He had milk every day." In a second experiment, subjects were timed as they read the completions. In both studies, context affected the ordering of exemplars. For the secretary example, tea was preferred to milk, while the reverse was true in the truck driver example. Neither result is consistent with the refocusing hypothesis.

2. *Differences in ideal points.* Another approach to changes in organization is to assume not that one shifts to a new best example but rather that the ideal point or prototype may change. Consider again the example from Roth and Shoben involving beverages. If one assumes that tea, coffee, and milk are arranged in a linear manner, then one might suggest that in the context of secretaries the ideal may be some point between tea and coffee but in the context of truck drivers the ideal beverage shifts to a point between coffee and milk. This notion of ideal points is in the spirit of Coomb's unfolding theory (Coombs, 1964). It appears to be consistent with the Roth and Shoben results but does not entail the assumption that the relations among concepts shift as contexts change. To apply this notion to our present results requires a set of assumptions for what drives changes in the ideal point. We see no principled means of doing this. Furthermore, if there is some context in which tea and milk are more similar than coffee and milk (e.g., at tea time), then one would be unable to capture the complete set of changes in terms of shifts in an ideal point.

3. *Differences in dimensional weighting.*One of the most common approaches to knowledge reorganization is to assume that different contexts lead some dimensions to be weighted more heavily than others. This assumption is the core idea for Tversky's striking demonstrations of diagnosticity effects and is reflected in multidimensional scaling programs that attempt to describe individual differences (e.g., the INDSCAL program of Carroll and Chang, 1970). Changes in dimension weights (e.g., on color and on monetary value) might well account for differences in similarity among gold, silver, and brass railings compared to gold, silver, and brass coins. It would not, however, account for the corresponding set of results for more unidimensional adjectives such as white, gray, and black clouds compared to hair, unless one were allowed free rein to posit sufficient dimensions so that, in effect, there were no constraints at all in the model. Again, one would also need to add processing assumptions concerning what drives changes in dimension weights and these processing assumptions would have to do the lion's share of the explanatory work.

4. *Local rescaling.* Changes in dimensional weights correspond to expanding or contracting a dimension uniformly. It is possible, of course, that attention to a subset of values on some dimension leads to local

changes in similarity relationships. Parducci's (1965) range-frequency theory and Krumhansl's (1978) suggestion that the local density of exemplars leads to a local increase in sensitivity seem consistent with this general principle. The local rescaling idea, although interesting, is an unlikely candidate for accounting for our present results. Although density of examples may influence similarity judgments it is not clear how to extend the density principle to any of our main experimental manipulations.

5. *Differences in ideals and in dimensional weights.* Obviously, one can combine two or more of the preceding principles for knowledge reorganization. At an abstract level, the Smith and Osherson modification model can be characterized as involving a change in the ideal point (via vote shifting) plus a change in dimensional weights (via a boost in diagnosticity). Barsalou's (1985) work on ideals as determinants of typicality also posits a model that falls into this class. Although these two models can certainly point to some successes, the present experiments suggest that they will prove inadequate as general models for conceptual combination.

Summary. The above set of models encompasses the most common approaches to knowledge reorganization. Since none of them can capture our three main results it appears that we will have to look elsewhere for viable models.

HYBRID MODELS

So far, the exemplar-based models of categorization have escaped our criticism. According to the exemplar view (Medin, 1986; Medin & Schaffer, 1978), categories are represented as a set of exemplars. Consequently, categorization decisions are based not on a comparison to prototypes, but instead on the retrieval of exemplars. In this view, combined categories may be thought of as limited sets of exemplars. Thus, for example, the concept of wooden spoon can be generated by deleting all the spoon exemplars that are not made of wood. Moreover, correlated attributes can be derived from this mechanism. One can determine that most wooden spoons are large simply by consulting the exemplars of this combined concept.

One immediate problem with this exclusionary rule is that it seems to imply that no concept that is not a member of the simple category can be a member of the conjoined category. Although this restriction seems to pose no problems for the concepts *spoon* and *wooden spoon*, Hampton (1987b) has demonstrated that there are categories for which this strict application of class inclusion will not capture people's judgments of category membership. For example, subjects agreed that blackboard was a relatively good example of the category *school furniture*. At the same

time, they also agreed that blackboards were not furniture. Without additional assumptions, an exclusionary rule cannot accommodate such a finding.

In addition, exemplar models in general tend to be short on processing details. Thus, for example, when one is asked if blackboards are furniture, is one's task to search the set of furniture exemplars or is it to compare a blackboard to each exemplar stored with furniture? If the latter, then is blackboard compared to all of these exemplars? How many are there? And what is the decision rule?

Finally, exemplar models do not incorporate causal knowledge nor do they ascribe a role for theories in organizing concepts. Thus they cannot account for the differences between judgments of white, gray, and black clouds versus hair or for our results on property centrality.

These problems lead us away from the strong claim that combined concepts are derived exclusively from selecting among stored examples. Instead, we propose that new representations may be derived as needed both from lists of exemplars and from other causal knowledge about the world. Thus, "found in a school" may greatly heighten an object's degree of membership in the category "school furniture." Similarity relations may be similarly affected. As we suggested earlier, the similarity of white and gray may be increased in the context of hair color because of our knowledge about the effects of age on hair color, and the similarity may be decreased in the context of cloud color because of our knowledge of the causal relationship between cloud color and the probability of precipitation.

Moreover, this knowledge of the world also enables us to determine what aspects of a concept are more essential, and therefore less changeable, than others. Kahneman and Miller (1986) have referred to this phenomenon as mutability. We have maintained that attributes that accept little change are relatively central to the concept's meaning and consequently we have used the term centrality rather than mutability. In any event, many of our subjects in our third experiment used counterfactuals, as described by Kahneman and Miller, to determine their ratings. Some noted, for example, that one might be able to obtain green swans through selective breeding.

Our results converge nicely with both arguments and evidence provided by Murphy (1987). He found that attributes listed for combined concepts were not a proper subset of attributes listed for the constituent concepts (e.g., "cooked in a pie" is a property of *sliced apples* but not of apples or sliced things in general) and that definitions of adjectives varied substantially across noun context. Murphy argues that these results require that world- and theory-based knowledge play a critical role in understanding combined concepts.

Although the principle of centrality and the importance of causal relations may point us in the right direction, we are a long way from a process model of conceptual combination that is at the level of detail of the modification model. We need to be able to specify the conceptual underpinning of our intuitions about centrality that seemed to underlie judgments in our third experiment. In seeking the answer to the puzzle of the origin of centrality, it seems to us that one might look to theory-based effects. For example, we may believe that a straight banana is still a banana because we have no theoretical belief that shape plays any major role in the definition of banana. On the other hand, a straight boomerang seems almost anomalous because we believe that it is the shape of the boomerang that causes it to return when thrown. Reference to Table 4 will demonstrate that centrality involves more than simple structure–function correlations. For example, structure–function correlations might lead one to expect that a soft knife would be at least as anomalous as a soft diamond. Softness in diamonds, however, would have important ramifications for a large body of knowledge in a way that a soft knife would not (presumably one could manufacture a soft knife).

Theoretical concerns may also play a role in comprehending categorical information in context. For example, given the sentence, "The banker's wife enjoyed wearing her fur coat to opening night at the opera" followed by "She became enamored with raccoon later in life," we may have to construct a scenario in order to understand this deviation from mink as the expected instantiation of fur coat. One possibility is that we may come to view the banker's wife as somewhat eccentric or nonconforming in order to explain her choice. That is, centrality may be driven, in part, by the types of explanations we construct in comprehending particular contexts.

Our data undermine a large class of current models, not just the modification model. It appears that concept modifications associated with context effects and conceptual combination require a dynamic restructuring of information. Obviously the problem of concept modification is extremely difficult but any insights gained in this complex domain may generate better theories of how simple concepts are structured to begin with. None of the main approaches to concepts (see Medin & Smith, 1984; Mervis & Rosch, 1981; Oden, 1987, for reviews; Smith & Medin, 1981), for example, contain a notion of centrality (but see Medin & Ortony, 1987). We think that centrality and the use of theoretical knowledge are necessary for a full understanding of how people use concepts. We may not need to solve the problem of conceptual combination in order to gain a complete understanding of simple concepts, but it seems to us that such an understanding is much more likely if we have some knowledge of how concepts are combined.

APPENDIX

Mean Typicality Ratings for the Various Adjective–Noun Combinations in the First Experiment

Adjective/noun	Average (12)	Adjective/noun	Average (13)
l = large, s = small (spoon)		w = wooden, m = metal (spoon)	
l as spoon	5.42	w as spoon	4.77
s as spoon	7.58	m as spoon	7.54
l as wooden spoon	7.25	w as large spoon	8.23
l as metal spoon	4.67	w as small spoon	2.54
s as wooden spoon	2.33	m as large spoon	4.69
s as metal spoon	7.75	m as small spoon	7.14
l = light colored d = dark colored (shirt)		s = summer shirt, w = winter shirt	
l as shirt	5.83	s as shirt	6.33
d as shirt	4.58	w as shirt	4.46
l as summer shirt	7.50	s as light-colored shirt	8.62
l as winter shirt	3.33	s as dark-colored shirt	2.23
d as summer shirt	2.50	w as light-colored shirt	3.23
d as winter shirt	8.08	w as dark-colored shirt	8.62
s = sour, S = sweet (fruit)		sm = small, l = large (fruit)	
s as fruit	4.08	sm as fruit	5.77
S as fruit	7.33	l as fruit	5.46
s as small fruit	5.25	sm as sour fruit	4.15
s as large fruit	4.42	sm as sweet fruit	5.08
S as small fruit	5.83	l as sour fruit	3.62
S as large fruit	5.42	l as sweet fruit	4.92
l = large, s = small (bird)		so = songless, si = singing (bird)	
l as bird	4.83	so as bird	2.69
s as bird	7.17	si as bird	7.92
l as songless bird	6.17	so as large bird	6.69
l as singing bird	2.83	so as small bird	2.69
s as songless bird	1.83	si as large bird	2.38
s as singing bird	7.00	si as small bird	7.23
c = clear, o = opaque (gem)		e = expensive, ch = cheap (gem)	
c as gem	7.08	e as gem	8.77
o as gem	5.58	ch as gem	2.69
c as expensive gem	6.67	e as clear gem	6.38
c as cheap gem	3.50	e as opaque gem	3.62
o as expensive gem	3.75	ch as clear gem	3.62
o as cheap gem	5.08	ch as opaque gem	5.00
r = round, R = rectangular (watch)		c = conventional, d = digital (watch)	
r as watch	7.42	c as watch	5.23
R as watch	5.00	d as watch	6.69
r as conventional watch	7.08	c as round watch	7.23
r as digital watch	2.58	c as rectangular watch	4.31

APPENDIX—*Continued*

Adjective/noun	Average (12)	Adjective/noun	Average (13)
R as conventional watch	3.42	d as round watch	3.31
R as digital watch	7.17	d as rectangular watch	7.85
s = small, l = large (boat)		s = sail, d = diesel engine (boat)	
s as boat	6.67	s as boat	7.46
l as boat	4.83	d as boat	4.46
s as sail boat	6.00	s as small boat	4.69
s as diesel engine boat	1.75	s as large boat	5.77
l as sail boat	4.50	d as small boat	2.85
l as diesel engine boat	6.08	d as large boat	7.92
F = Florida, M = Minnesota (roof)		f = flat, s = slanted (roof)	
F as roof	2.92	f as roof	4.85
M as roof	4.67	s as roof	7.23
F as flat roof	6.00	f as Florida roof	4.38
F as slanted roof	2.83	f as Minnesota roof	2.15
M as flat roof	1.17	s as Florida roof	4.38
M as slanted roof	5.00	s as Minnesota roof	5.38
l = lace, f = flannel (nightgown)		b = black, p = pink (nightgown)	
l as nightgown	6.75	b as nightgown	4.85
f as nightgown	4.25	p as nightgown	5.62
l as black nightgown	7.17	b as lace nightgown	6.85
l as pink nightgown	4.17	b as flannel nightgown	1.69
f as black nightgown	2.00	p as lace nightgown	6.69
f as pink nightgown	3.58	p as flannel nightgown	3.85
c = cultivated, w = wild (flower)		l = large, s = small (flower)	
c as flower	5.83	l as flower	5.08
w as flower	5.75	s as flower	6.00
c as large flower	4.42	l as cultivated flower	5.46
c as small flower	4.17	l as wild flower	4.15
w as large flower	3.92	s as cultivated flower	5.08
w as small flower	6.58	s as wild flower	7.15
b = black & white, c = color (TV)		s = small, l = large (TV)	
b as TV	3.42	s as TV	5.15
c as TV	8.25	l as TV	5.61
b as small TV	6.33	s as black & white TV	7.69
b as large TV	2.75	s as color TV	3.85
c as small TV	4.25	l as black & white TV	2.85
c as large TV	7.67	l as color TV	8.46
p = paperback, h = hardcover (book)		f = fiction, t = text (book)	
p as book	6.25	f as book	6.23
h as book	7.33	t as book	7.00
p as fiction book	7.08	f as paperback book	7.77
p as textbook	2.75	f as hardcover book	4.07
h as fiction book	4.58	t as paperback book	2.54
h as textbook	8.33	t as hardcover book	9.31

APPENDIX—*Continued*

Adjective/noun	Average (12)	Adjective/noun	Average (13)
h = harmless, t = threatening (cloud)		w = white, g = gray (cloud)	
h as cloud	6.25	w as cloud	8.38
t as cloud	4.25	g as cloud	6.23
h as white cloud	8.42	w as harmless cloud	9.08
h as gray cloud	1.92	w as threatening cloud	1.00
t as white cloud	0.92	g as harmless cloud	1.92
t as gray cloud	8.00	g as threatening cloud	8.92
s = short, l = long (grass)		g = green, b = brown (grass)	
s as grass	7.67	g as grass	9.46
l as grass	6.25	b as grass	2.08
s as green grass	7.25	g as short grass	5.92
s as brown grass	3.83	g as long grass	6.23
l as green grass	6.42	b as short grass	6.61
l as brown grass	3.67	b as long grass	3.31
j = juicy, d = dry (tomato)		s = summer, winter = winter (tomato)	
j as tomato	8.17	s as tomato	6.15
d as tomato	2.92	w as tomato	2.77
j as summer tomato	7.92	s as juicy tomato	7.46
j as winter tomato	3.42	s as dry tomato	2.38
d as summer tomato	2.00	w as juicy tomato	3.31
d as winter tomato	4.50	w as dry tomato	5.23
l = large, s = small (dog)		f = ferocious, t = timid (dog)	
l as dog	5.25	f as dog	5.23
s as dog	6.00	t as dog	3.85
l as ferocious dog	7.25	f as large dog	7.54
l as timid dog	2.00	f as small dog	3.54
s as ferocious dog	3.08	t as large dog	2.92
s as timid dog	6.25	t as small dog	6.08
p = paved, g = gravel (street)		b = busy, e = empty (street)	
p as street	8.58	b as street	7.85
g as street	4.00	e as street	4.15
p as busy street	7.67	b as paved street	8.92
p as empty street	2.75	b as gravel street	1.15
g as busy street	0.58	e as paved street	4.85
g as empty street	7.00	e as gravel street	7.85
s = small, l = large (ball)		h = hard, s = soft (ball)	
s as ball	6.33	h as ball	7.08
l as ball	5.58	s as ball	6.85
s as hard ball	7.25	h as small ball	7.54
s as soft ball	2.42	h as large ball	4.00
l as hard ball	2.75	s as small ball	3.63
l as soft ball	7.33	s as large ball	7.62
h = health, j = junk (food)		b = bland, t = tasty (food)	
h as food	4.08	b as food	4.15

APPENDIX—*Continued*

Adjective/noun	Average (12)	Adjective/noun	Average (13)
j as food	5.00	t as food	7.46
h as bland food	8.42	b as health food	7.00
h as tasty food	2.50	b as junk food	1.15
j as bland food	2.67	t as health food	4.08
j as tasty food	7.58	t as junk food	7.92
s = small, l = large (needle)		sh = sharp, d = dull (needle)	
s as needle	7.17	sh as needle	9.00
l as needle	5.08	d as needle	2.69
s as sharp needle	7.25	sh as small needle	7.23
s as dull needle	2.17	sh as large needle	4.62
l as sharp needle	4.25	d as small needle	2.08
l as dull needle	5.00	d as large needle	3.69

REFERENCES

Anderson, R. C., & Shifrin, Z. (1980). The meaning of words in context. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 331–348). Hillsdale, NJ: Erlbaum Associates.

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

Barsalou, L. W., & Sewell, D. R. (1984). *Constructing representations of categories from different points of view*. Emory Cognition Project Tech. Rep. No. 2, Emory University.

Carroll, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an *N*-way generalization of "Eckard–Young" decomposition. *Psychometrika*, **35**, 283–319.

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

Coombs, C. H. (1964). *A theory of data*. New York: Wiley.

Estes, W. K. (1986). Array models for category learning. *Cognitive Psychology*, **18**, 500–549.

Hampton, J. A. (1987a). Inheritance of attributes in natural concept conjunctions. *Memory & Cognition*, **15**, 55–71.

Hampton, J. A. (1987b). Overextension of conjunctive concepts: Evidence for a unitary model of concept typicality and class inclusion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, in press.

Jones, G. V. (1982). Stacks not fuzzy sets: An ordinal basis for prototype theory of concepts. *Cognition*, **12**, 281–290.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, **93**, 136–153.

Krumhansl, C. L. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review*, **85**, 445–463.

Levi, J. N. (1978). *The syntax and semantics of complex nominals*. New York: Academic Press.

Malt, B. C., & Smith, E. E. (1983). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior*, **23**, 250–269.

Medin, D. L. (1986). Comment on "Memory Storage and Retrieval Processes in Category Learning." *Journal of Experimental Psychology: General*, **115**, 373–381.

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., & Ortony, A. (1987). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning*. New York: Cambridge University, in press.

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

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

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, concept cohesiveness, and category construction. *Cognitive Psychology*, **19**, 242–279.

Mervis, C. B., & Roth, E. M. (1981). The internal structure of basic and nonbasic color categories. *Language*, **57**, 384–405.

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

Murphy, G. L. (1987). Comprehending complex concepts. *Cognitive Science*, in press.

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

Oden, G. C. (1984). *Everything is a good example of something, and other endorsements of the adequacy of a fuzzy theory of concepts*. WHIPP 21, Wisconsin Human Information Processing Program, Madison, WI.

Oden, G. C. (1987). Concept, knowledge, and thought. *Annual Review of Psychology*, **38**, 203–227.

Osherson, D. N., & Smith, E. E. (1982). Gradedness and conceptual combination. *Cognition*, **12**, 299–318.

Parducci, A. (1965). Category judgment: A range–frequency model. *Psychological Review*, 72, 407–418.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, **104**, 192–223.

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

Smith, E. E. (1987). Concepts and thought. In R. J. Sternberg & E. E. Smith (Eds.), *The psychology of human thought*. Cambridge, MA: Cambridge Univ. Press.

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

Smith, E. E., & Osherson, D. N. (1984). Conceptual combination with prototype concepts. *Cognitive Science*, **8**, 337–361.

Smith, E. E., & Osherson, D. N. (1987). Compositionality and typicality. In S. Schifter & S. Steele (Eds.), *The 2nd Arizona Colloquium on Cognitive Science*. Tucson, AZ: Univ. of Tucson Press.

Smith, E. E., Osherson, D. N., Rips, L. J., Albert, K., & Keane, M. (1986). Combining prototypes: A modification model. Unpublished manuscript.

Thagard, P. (1984). Conceptual combination and scientific discovery. In P. Asquith & P. Kitcher (Eds.), *PSA* (Vol. 1). East Lansing, MI: Philosophy of Science Association.

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

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.

Wyer R. S., & Srull, T. K. (1986). Human cognition in its social context. *Psychological Review*, **93**, 322–359.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, **8**, 338–353.

Zadeh, L. A. (1982). A note on prototype theory and fuzzy sets. *Cognition*, **12**, 291–297.

(Accepted September 10, 1987)