COGNITIVE PSYCHOLOGY 25, 231–280 (1993)

Feature-Based Induction

STEVEN A. SLOMAN

University of Michigan

A connectionist model of argument strength that applies to categorical arguments involving natural categories and predicates about which subjects have few prior beliefs is proposed. An example is *robins have sesamoid bones, therefore falcons have sesamoid bones*. The model is based on the hypothesis that argument strength is related to the proportion of the conclusion category's features that are shared by the premise categories. The model assumes a two-stage process. First, premises are encoded by connecting the features of premise categories to the predicate. Second, conclusions are tested by examining the degree of activation of the predicate upon presentation of the features of the conclusion category. The model accounts for 13 qualitative phenomena and shows close quantitative fits to several sets of argument strength ratings.

FEATURE-BASED INDUCTION

One way we learn about and function in the world is by inducing properties of one category from another. Our knowledge that leopards can be dangerous leads us to keep a safe distance from jaguars. Osherson, Smith, Wilkie, Lopez, and Shafir (1990) examine the conditions under which a property that is asserted of one or more categories will also be asserted of another category by asking about the judged strength of *categorical* arguments such as

Elephants love onions.
Mustangs love onions.
———————————————— (a)
Therefore, Zebras love onions.

What degree of belief in the conclusion of such an argument is attributable to the premises and what is the nature of the inductive inference?

Osherson et al. (1990) propose that the psychological strength of categorical arguments depends on "(i) the degree to which the premise categories are similar to the conclusion category and (ii) the degree to which

This research was conducted while the author held a postdoctoral fellowship from the Natural Sciences and Engineering Research Council of Canada. I thank James Morgan for assisting with data collection and Ed Wisniewski, Keith Holyoak, Dan Osherson, Doug Medin, John Kruschke, two anonymous reviewers, and especially Ed Smith for helpful comments and ideas. Correspondence should be addressed to the author who is now at the Department of Cognitive and Linguistic Sciencies, Box 1978, Brown University, Providence, RI 02912.

0010-0285/93 $9.00

the premise categories are similar to members of the lowest level category that includes both the premise and the conclusion categories" (p. 185). Their model accounts for 13 qualitative phenomena and provides good quantitative fits to the results of several experiments (corroborative data are presented in Osherson, Stern, Wilkie, Stob, & Smith, 1991, and Smith, Lopez, & Osherson, 1992).

This paper proposes an alternative, feature-based, model of this inductive task. The model is expressed using some simple connectionist tools, similar to those that have been used to model learning processes (e.g., Gluck & Bower, 1988). The main idea is that an argument whose conclusion claims a relation between category C (e.g., Zebras) and predicate P (e.g., love onions) is judged strong to the extent that the features of C have already been associated with P in the premises. The automatic generalization property of distributed representations is exploited in order to model the induction of predicates from one category to another (cf. Hinton, McClelland, & Rumelhart, 1986). The model accounts for 10 of the phenomena described by Osherson et al. (1990) and provides a more accurate account of one of the remaining phenomena. A generalization of the model, discussed near the end of the paper, is shown to be compatible with the remaining 2 phenomena. The model also motivated 2 new phenomena. Finally, the model is shown to provide good fits to argument strength ratings.

Osherson et al.'s model is *category-based* in that it assumes that judgments of argument strength depend on a stable hierarchical category structure that is describable without reference to the attributes or features of either category, such as the superset–subset relation that exists between Mammals and Elephants. Indeed, the primitive elements in Osherson et al.'s (1990) model are pairwise similarities between categories at the same hierarchical level. In Osherson et al. (1991), pairwise similarities are derived from feature vectors. Nevertheless, their model remains category-based in that features are used only to derive similarities between categories at the same hierarchical level.

The present model is *feature-based* in that it assumes that argument strength judgments depend on connection strengths between features of the conclusion category and the property of interest without regard to any fixed structural relations that may exist between whole categories. All categories are represented as vectors of numerical values over a set of features. The existence of a stable category-structure is not assumed by the model because it is not necessary; all inductive processes depend strictly on the features of premise and conclusion categories. Obviously, people do have some knowledge about the hierarchical organization among some categories. Many people know that Elephants are Mammals. The assumption being made here is that this knowledge is represented in

a way distinct from the structures that normally support judgments of categorical argument strength. Knowledge about category structure is not generally used when engaging in the kind of inductive task under consideration, although surely it is used some of the time.

The foregoing assumes a distinction between what might be called *intuitive* and *logical* modes of inference (cf. Rips, 1990, for a parallel distinction between the loose and strict views of reasoning). This kind of distinction has been implicit in much previous work on the psychology of reasoning. When, for example, Kahneman and Tversky (1973) propose that people employ a representativeness heuristic when making probability judgments, they do not exclude the possibility that people can also employ probability theory to make the same judgments. Judgments by representativeness are for the most part intuitive, whereas the explicit application of probability theory entails some logical thought (at least for nonexperts). Sometimes the two forms of reasoning lead to different conclusions (e.g., Tversky & Kahneman, 1983). I cannot provide strict criteria to discriminate occasions on which people will use one or the other kind of reasoning, but some general guidelines do seem apparent. People are more likely to use intuition when they lack relevant skill or knowledge (especially about causal relations), when they have not been trained, when they are short of time or will, when they feel no need to justify their response, and when they have not been instructed to do otherwise. The feature-based model is intended to capture an aspect of only the "looser," intuitive form of reasoning. As a model of argument strength, it presupposes that, given a fixed featural representation of each category, category hierarchies are needed to describe only the logical reasoning process, not the intuitive one. The similarity-coverage model can be construed as a model of intuition only if one is willing to ascribe categorical reasoning to an intuitive process.

A second difference between the two models is that only the category-based model assumes that subjects explicitly compute similarity. The feature-based model does assume a feature-matching process, but not one that computes any empirically valid measure of similarity per se.

The paper proceeds as follows. The Osherson et al. (1990) model is briefly described, followed by an introduction to the feature-based model. Next, the feature-based model's ability to account for each of the phenomena described by Osherson et al. (1990), and some new ones, is discussed. Osherson et al. (1990) point out that "the phenomena should . . . be conceived as tendencies rather than strict laws determining confirmation." Because neither model expects all of the phenomena to hold for every applicable argument, and because the large number of phenomena limits the number of arguments that each phenomenon has been tested on, much of the empirical support for the models rests on their

ability to quantitatively fit subjects' ratings of argument strength. The penultimate section of the paper demonstrates that the current model provides close fits to such data. Finally, I discuss the strengths and weaknesses of the two models and possible extensions of the feature-based one.

The Similarity-Coverage Model of Argument Strength

My presentation of the category-based model will be terse, since my aim is to focus on the feature-based model. Osherson et al. (1990) develop their model using an extended pairwise similarity function, SIM_S, defined for a given subject S. They suppose the existence of $SIM_S(k;g)$ which returns a real number between 0 and 1 reflecting the similarity between any pair of elements k and g that are at the same hierarchical level within some natural category. Osherson et al. (1990) extend SIM_S in two ways. First, they treat multiple-premise arguments by employing a notion of similarity that obtains between several category members, $k_1 \ldots k_n$, and another category member, g, all of which reside at the same hierarchical level. $SIM_S(k_1 \ldots k_n;g)$ is defined as $MAX\{SIM_S(k_1;g), \ldots, SIM_S(k_n;g)\}$. In other words, the similarity between $k_1 \ldots k_n$ and g is defined as the maximum of the n pairwise similarities between each of the k's and g. The MAX rule is important because it defines similarity using a nearest-neighbor principle. As will become clear below, it implies that an argument can be strong if the conclusion category is similar to only a single premise category.

Second, Osherson et al. (1990) extend SIM_S to treat cases in which categories are not all at the same hierarchical level. If G is a category at a higher level than $k_1 \ldots k_n$, $SIM_S(k_1 \ldots k_n;G)$ is defined as the average of

$$\{SIM_S(k_1 \ldots k_n;g) | \text{S believes that } g \text{ is at the same level as } k_1 \ldots k_n \text{ and that } g \text{ belongs to } G\}.$$

"$SIM_S(k_1 \ldots k_n;G)$ is the average similarity that S perceives between $k_1 \ldots k_n$ and members of G at the level of $k_1 \ldots k_n$" (Osherson et al., 1990, p. 191).

Both the category and feature-based models of argument strength apply to arguments, such as (a) above, that can be written schematically as a list of sentences, $P_1 \ldots P_n/C$, in which the P_i are the premises of an argument and C is the conclusion, each with an associated category (cat). Osherson et al.'s (1990) model consists of a linear combination of two variables defined in terms of SIM_S. The first variable, similarity, expresses the degree of resemblance between premise categories and the conclusion category. The second variable, coverage, reflects the degree of resemblance between premise categories and members of the lowest-level cat-

egory that properly includes both the premise and conclusion categories. Therefore, they refer to their model as the similarity-coverage model of argument strength.

Denote [cat(P_1), . . . , cat(P_n), cat(C)] as the lowest level category K such that the category in each of P_1, . . . , P_n and C is a subset of K. For example, [Elephant, Zebra] = Mammal. The formal statement of the similarity-coverage model is that for every person S there is a positive constant $\beta \in (0,1)$ such that for all arguments A = P_1 . . . P_n/C, the strength of A for S is given by

$$\beta SIM_S(cat(P_1), \ldots, cat(P_n); cat(C)) + (1 - \beta)SIM_S(cat(P_1), \ldots, cat(P_n); [cat(P_1), \ldots, cat(P_n), cat(C)].$$

The arguments examined by Osherson et al. (1990) were all categorical in that premises and conclusion all had the form "all members of Y have property X" where Y was a simple category (like Feline or Reptile) and X remained fixed across the premises and conclusion within each argument (cf. Rips, 1975). Categories all involved natural kinds, in particular the hierarchy of living things. Predicates were chosen so that subjects would have few prior beliefs about them, for instance "secretes uric acid crystals." Osherson et al. (1990) dub these *blank* predicates, and point out that their use permits theorists to focus on the role of categories in the transmission of belief from premises to conclusion, minimizing the extent to which subjects reason about the particular properties employed. Some comments and speculation about modeling nonblank predicates can be found in this paper's concluding section. Otherwise, my discussion is limited to blank predicates whose identity is irrelevant and therefore will be generically referred to as "predicate X."

Osherson et al. (1990) showed that their model accounts for 13 qualitative phenomena—described in detail below—and used similarity ratings to generate predictions from their model that were highly correlated with subjects' ratings of argument strength in several experiments (see also Smith et al., 1992, and Osherson et al., 1991), some of which will also be described below. I now offer my alternative feature-based model.

THE FEATURE-BASED MODEL

Overview

We start with a set of input units to encode feature values and an output unit to encode the blank predicate X. Consider the argument

$$\frac{\text{Robins have } X.}{\text{Falcons have } X.} \quad \text{(b)}$$

The process by which the feature-based model determines the strength

of this argument is diagramed in Fig. 1. The state of the network before presentation of the argument is illustrated in Fig. 1a. Because the argument's predicate is blank, the unit representing it is initially not connected to any featural units. Premises are then encoded by connecting the units representing the features of each premise category to the predicate unit allowing the category units to activate the predicate unit (Fig. 1b). To encode the premise of Argument (b) for instance, the units encoding features of Robins (a small number of binary features are used for expository purposes only) are connected to the blank predicate unit X. Had there been more than one premise in Argument (b), the features of each additional premise's category would have been connected to unit X in identical fashion. Second, conclusions are tested by determining the extent to which unit X becomes "activated" upon presentation of the features of the conclusion category (Fig. 1c). In the example, this would be accomplished by observing the activation value of unit X upon presentation of the features of Falcons which would be 0.5 because the predicate unit is connected to one-half of the active category units.

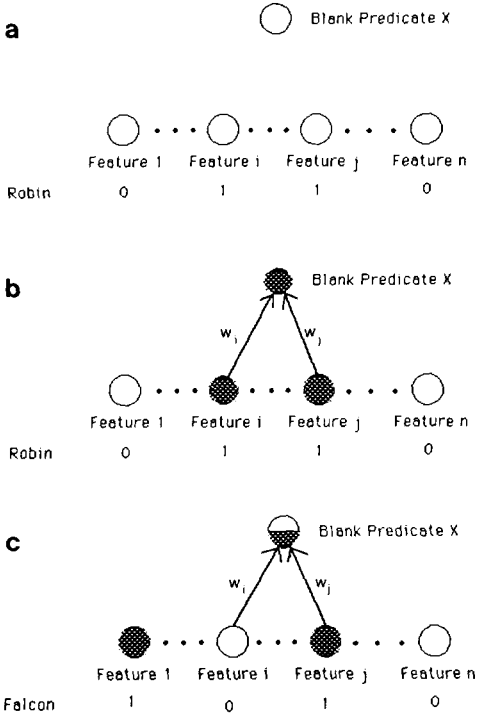

FIG. 1. Illustration of the feature-based model for the argument "Robins have property X, therefore Falcons have property X." (a) Before encoding premise "Robins have X." (b) After encoding premise "Robins have X." (c) Testing conclusion "Falcons have X."

I posit two rules, an encoding rule and an activation rule, which, along with a set of feature vectors, completely determine the model's predictions. The encoding rule, Eq. 1 below, defines how connections are established. The activation rule, Eq. 3 below, determines the value of unit X (denoted a_x). Because connection strengths depend on the values of premise features, a_x increases with the extent of shared features between premise and conclusion categories. a_x is defined to decrease with the extent of features in the conclusion category alone.

Throughout the paper, vectors will be represented by capital letters (e.g., vector B) and their scalar elements will be represented by corresponding lowercase letters subscripted by their ordinal position (e.g., b_1 or b_i).

Categories and Predicates

Every category, such as Robins, Mammals, or Animals, is identified with a vector F of n real numbers from the closed [0,1] interval. For instance, F(Robins) = $[f_1(\text{Robins}), \ldots, f_n(\text{Robins})]$, where F(Robins) is a vector encoding the feature values of Robins. The f_i are variables which represent a basis set of atomic features and are assigned individual feature values, in this case those corresponding to Robins. Both the vector F and the scalar f_i's could be subscripted by S, to indicate that F refers to a given subject S's representation of a category, because a category's representation presumably varies from subject to subject. We could also index representations by time. Such indices will be left implicit in what follows.

Feature values used for modeling can be derived by having subjects list or rate features or attributes for each category. I do not assume that vectors represent sets of necessary or sufficient features, nor that features are independent of each other. I rather assume that they represent a large number of interdependent perceptual and abstract attributes. In general, these values may depend on the context in which categories are presented. I will assume however that category representations are fairly constant within each of the experiments to be discussed. Notice that no distinction is made between the nature of the representations of general categories like Animals and specific categories like Robins.

Blank predicates will be identified with a single scalar variable. Because predicates can be semantically rich themselves, we may be tempted to represent them as vectors as well. At this point, however, because all predicates are blank, such a treatment would only serve to introduce an unnecessary level of complexity. All the demonstrations below could be altered to allow for a vector representation of predicates.

Network Architecture and Dynamics

The model of any single argument uses a network of n input units and

a single output unit. Each input unit is assigned value a_i equal to f_i, the value of feature i of the category under consideration. So, upon presentation of category Y, activation vector $A(Y) = [a_1(Y), \ldots, a_n(Y)] = F(Y)$; the current stimulus $A(Y)$ is set equal to the stored representation $F(Y)$. The activation of the output unit, a_x, represents the willingness to confirm predicate X of the input category. Upon presentation of a conclusion category, if the value of a_x is high, then the argument is judged strong; if low, then it is judged weak. Finally, connecting input units to predicate unit X is a vector of weights, $W = [w_1, \ldots, w_n]$. All weight values are restricted to the closed interval [0,1].

A, a_x, and W are dynamic variables and should somehow be indexed by their history. A is already indexed by the current input. W depends only on encoded premises P_1 to P_j. It will be indexed as $W(P_1, \ldots, P_j)$. When encoding the jth premise in a multiple-premise argument, a_x depends on encoded premises P_1 to P_{j-1} and input premise P_j, in which case I will write $a_x(P_j/P_1, \ldots, P_{j-1})$. When testing a conclusion, a_x depends on encoded premises P_1 to P_j and conclusion C, and I will write $a_x(C/P_1, \ldots, P_j)$. I will compress the notation by not writing out predicates because they are blank and therefore uninformative. For example, the strength of Argument (a) would be written a_x(Zebras/Elephants, Mustangs). If no premises have previously been encoded, the value of a_x given premise P as input will be denoted $a_x(P)$.

Encoding a Premise (Learning Rule)

To encode the category in a premise P, input units are set equal to the feature values of the category in P, so that $A(P) = F(P)$, and then weights are changed according to the following delta rule (cf., Sutton & Barto, 1981) in which the network is learning to turn on unit X in the presence of input from the category of P. Let the weight vector have some value $W(P_0)$ where P_0 represents zero or more premises, then after encoding premise P,

$$w_i(P_0,P) = w_i(P_0) + \lambda_i[1 - a_x(P/P_0)]a_i(P), \quad (1)$$

where each λ_i is a scalar coefficient. To keep each weight between 0 and 1, I assume that a_x does not exceed 1 and set

$$\lambda_i = 1 - w_i(P_0). \quad (2)$$

This removes the one free parameter we otherwise would have had. It also ensures that weights never decrease.

An important property of this encoding rule is that it depends on the activation of unit X. The extent to which each premise category activates unit X is established before that premise is encoded. Encoding a premise involves updating connection strengths. The amount of change that con-

nection strengths undergo is proportional to the premise's *surprise* value $[1 - a_x(P/P_0)]$, or the extent to which the premise category does not activate unit X given the premises already encoded.

Below, we will see that the strength of an argument in which the premise and conclusion categories are the same is 1; given some premise Y, $a_x(Y/Y) = 1$. Because weight change is proportional to $1 - a_x$, a corollary is that no weight change obtains after the first presentation of a repeated premise. In other words, the model predicts that an argument with multiple presentations of the same premise will be equally as strong as an argument with only a single presentation. The failure of such a prediction would be startling and presumably due to some sort of pragmatic violation.

Activation Rule for Unit X

The activation of unit X depends on the premises P_1 to P_j that have already been encoded in the weights as well as the current input category I which comes either from another premise or from the conclusion. The activation rule is

$$a_x(I/P_1, \ldots, P_j) = \frac{W(P_1, \ldots, P_j) \cdot A(I)}{|A(I)|^2}, \tag{3}$$

where $\cdot$ means the dot or inner product, defined as $U \cdot V = \sum_{i=1}^{n} u_i v_i$, which is a measure of the two vectors' overlap. The vertical bars represent the length of a vector. By "magnitude," I refer to the denominator of (3), or length squared: $|U|^2 = \sum_{i=1}^{n} u_i^2$. In words, the activation of unit X is proportional to the overlap between values of corresponding weights and input elements and inversely proportional to the number and size of input elements. When I is a conclusion category, $a_x(I/P_1, \ldots, P_j)$ is a model of argument strength.

Because weights and activations are always greater than or equal to 0, so is a_x. Under certain unusual circumstances, a_x may exceed 1. This will occur if, for instance, the magnitude of the conclusion category vector is small (see also the discussion of phenomenon *x*, premise–conclusion identity, below). In order to ensure a maximum value for the strength of an argument, and to maintain the property that weights never decrease, the activation rule could be extended to cut the value of a_x off at 1. This minor extension of the model has not yet proven necessary and will therefore not be implemented.

Psychological Interpretation of Activation Rule

Equation 3 states that the activation value of unit X increases with the dot product of the current input and previous inputs as they are encoded

in the weights by the learning rule. The interpretation is that a novel predicate is affirmed of the current input category to the extent that it shares features with other categories which are known to have the relevant property. If we know that Robins have some property X, then to the extent that Robins and Falcons have other properties in common, we will judge that Falcons also have property X.

Equation 3 also states that the activation value of unit X is inversely proportional to the magnitude of the current input. Magnitude is a measure of the number and size of the features in a category's representation. It is meant to capture the richness of a representation by indicating the extent of salient features in the category. The magnitude of a representation would be determined by the category's familiarity and complexity. The claim is that our willingness to affirm a property of a category decreases with the amount we already know about that category, given that the number of features the category shares with other categories that possess that property is held constant.

The most straightforward test of the conclusion magnitude prediction requires a specification of the features of some set of categories so that we can measure their common features and magnitudes directly. Throughout this paper, I will make use of feature ratings collected by Osherson et al. (1991), who obtained ratings of "the relative strength of association" between a set of properties and a set of mammals.[1] These ratings may not be ideal for testing the feature-based model. A more appropriate judgment task would instead have asked the subject about the relative prominence of each property, or the extent to which the subject attends to property X when thinking about Mammal Y. However, because these ratings are obviously estimates of people's knowledge about the relevant categories and the model requires no detailed analysis of the features themselves in any case, and because they were carefully collected and are clearly not biased in favor of the feature-based model, they will serve us adequately here.

I constructed pairs of one-premise arguments by choosing triples of Mammal categories such that one Mammal shared an equal number of features with both of the other categories but the two other categories differed in their magnitudes. For example, according to Osherson's feature-ratings, Collies and Horses have about as many features in common as Collies and Persian cats; i.e., $F(\text{Collies}) \cdot F(\text{Horses}) \approx F(\text{Collies}) \cdot F(\text{Persian cats})$. But the representation of Horses is of greater

[1] I thank Daniel Osherson and Tony Wilkie for making these feature ratings available. Subjects rated 48 mammals for 85 properties on a scale that started from 0 and had no upperbound. Ratings were provided by 8 or 9 M.I.T. students for each mammal. The ratings were linearly transformed to values in the [0,1] interval and then averaged.

magnitude than that of Persian cats. $|F(\text{Horses})|^2 > |F(\text{Persian cats})|^2$. Consider the pair of arguments

$$\frac{\text{All Collies produce phagocytes.}}{\text{All Persian cats produce phagocytes.}} \quad \text{(c)}$$

and

$$\frac{\text{All Collies produce phagocytes.}}{\text{All Horses produce phagocytes.}} \quad \text{(d)}$$

According to the feature-based model, the weight vectors obtained by encoding the premises of each argument will be identical because the premises themselves are identical. The premise and conclusion categories in the two arguments have the same measure of common features, so any difference in strength between the two arguments must be attributable to the difference in magnitude of their conclusion categories (this is further explicated below, by Eq. 5). In particular, people should find Argument (c) stronger than Argument (d) because it has a lower magnitude conclusion category.

I asked 34 University of Michigan undergraduates to rate the convincingness of each of 20 arguments, 10 pairs of arguments that conformed to this structure. Within a pair, premises were identical and shared an equal measure of common features with the two conclusion categories (the mean dot product between premise and larger magnitude conclusion was 5.31, essentially identical to the value for the smaller magnitude conclusion of 5.32). But conclusions had different magnitudes (the mean of the larger magnitudes was 10.74, the mean for the smaller was 7.66). Subjects rated each argument on an interval scale in which 0 meant very unconvincing and 10 meant very convincing (details of the experimental design and procedure can be found in Appendix A).

For each pair, the argument with the smaller magnitude conclusion category was rated as stronger (mean rating was 2.17) than that with the larger magnitude conclusion category (mean of 1.65). This difference was highly significant for these 10 argument pairs across subjects, $t(33) = 2.75$, $p < .01$. The feature-based model has successfully predicted a rather nonintuitive result: The strength of an argument can be increased by choosing a conclusion category which has fewer features associated with it, even when a measure of features common to the premise and conclusion categories is held constant.

The magnitude prediction also finds support in people's tendency to generalize less from a member of an in-group to other in-group members than from a member of an out-group to other out-group members (Quattrone & Jones, 1980). For example, Rutgers sophomores who were told

that another Rutgers student had chosen to listen to classical music over rock music gave lower percentage estimates that other Rutgers students would also choose classical music relative to estimates they gave that Princeton students would choose classical music after observing a Princeton student do so. This tendency supports the current hypothesis since people presumably know more about the category of people in their in-group than in an out-group.

Feature Coverage

According to the model, argument strength is, roughly, the proportion of features in the conclusion category that are also in the premise categories. For single-premise arguments in which features are all either 0 or 1, this statement is exact. An exact formulation for the general case is provided by the geometric interpretation below. Intuitively, an argument seems strong to the extent that premise category features "cover" the features of the conclusion category, although the present notion of coverage is substantially different from that embodied by the similarity-coverage model.

Single-Premise Arguments

Before encountering an argument, weights are all equal to 0, indicating that the network has no prior knowledge relevant to the property; i.e., the property is blank. In other words, $w_i() = 0$ for all i, where the empty parentheses indicate that nothing has been encoded. Therefore, using Eq. 3, the value of a_x as the first premise is encoded is

$$\begin{aligned} a_x(\mathrm{P}_1) &= \frac{\mathrm{W}() \cdot \mathrm{A}(\mathrm{P}_1)}{|\mathrm{A}(\mathrm{P}_1)|^2} \\ &= \frac{\Sigma w_i() a_i(\mathrm{P}_1)}{\Sigma a_i(\mathrm{P}_1)^2} \\ &= 0, \end{aligned}$$

and the value of each weight after encoding a single premise is

$$\begin{aligned} w_i(\mathrm{P}_1) &= w_i() + [1 - w_i()][1 - a_x(\mathrm{P}_1)]a_i(\mathrm{P}_1) \\ &= 0 + (1 - 0)(1 - 0)a_i(\mathrm{P}_1) \\ &= a_i(\mathrm{P}_1) \\ &= f_i(\mathrm{P}_1). \end{aligned} \tag{4}$$

Therefore, the weight vector after encoding a single premise is identical to the vector representing the category in the premise, $\mathrm{W}(\mathrm{P}_1) = \mathrm{F}(\mathrm{P}_1)$. Furthermore, the strength of an argument consisting of a single premise and a conclusion is, using first (3) and then (4):

$$a_x(C/P_1) = \frac{W(P_1) \cdot F(C)}{|F(C)|^2} = \frac{F(P_1) \cdot F(C)}{|F(C)|^2}. \quad (5)$$

In sum, single-premise arguments depend only on the dot product of the vectors representing premise and conclusion categories and, inversely, on the magnitude of the conclusion category vector.

Geometric Interpretation

By considering a geometric representation of the feature-based model, we can state precisely the sense in which it posits that the strength of an argument is equal to the proportion of the conclusion category's features that it shares with the premise categories. We require the notion of a projection of vector W on vector F ($Proj_F W$). $Proj_F W$ is a vector that can be thought as the "F-component" of W (see Fig. 2a). It is defined as

$$Proj_F W = \left[\frac{W \cdot F}{|F|^2}\right] F.$$

$[(W \cdot F)/|F|^2]$ is a scalar so that $Proj_F W$ is a vector with the same direction as F that can differ in length. Given an argument with premises $P_1, \ldots, P_j$, j applications of the encoding rule will produce weight vector

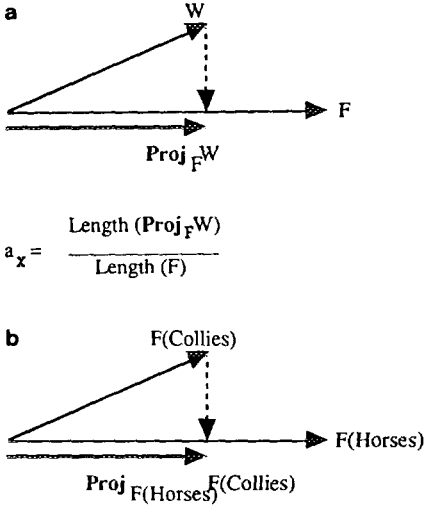

FIG. 2. Geometric interpretation of the feature-based model in terms of projections. (a) General case. (b) Illustration of single-premise argument.

$W(P_1, \ldots, P_j)$. Given conclusion C and conclusion category vector F(C), Eq. 3 tells us that the strength of the argument is

$$a_x(C/P_1, \ldots, P_j) = \frac{W(P_1, \ldots, P_j) \cdot F(C)}{|F(C)|^2}.$$

The weights are a (nonlinearly derived) representation of the premises. Therefore, the projection of the weights onto the conclusion category vector is a representation of the featural components of the premises that are also in the conclusion. It corresponds to those features that are shared by the conclusion category and any of the premises as they are encoded in the weights. Substituting argument strength into the definition of the projection shows that

$$\text{Proj}_{F(C)} W(P_1, \ldots, P_j) = a_x(C/P_1, \ldots, P_j)\, F(C).$$

By calculating the lengths of the two vectors (one on each side of the equation) and rearranging, we see that

$$a_x(C/P_1, \ldots, P_j) = \frac{|\text{Proj}_{F(C)} W(P_1, \ldots, P_j)|}{|F(C)|}.$$

In words, the strength of an argument is equal to the proportion of the length of the conclusion category vector that is spanned by the length of the projection of the weight vector onto the conclusion category vector. In this sense, argument strength is the proportion of features in a conclusion category that are also in the premise categories.

In the previous section, we saw that the weight vector used to determine the strength of a single-premise argument is identical to the vector representing the premise category. Therefore, as depicted in Fig. 2b, the strength of a single-premise argument is simply the proportion of the length of the conclusion category vector that is spanned by the length of the projection of the premise category vector onto the conclusion category vector. One interpretation of Fig. 2b is that an argument is strong to the extent its conclusion can be "justified" or "explained" by its premises relative to the total amount that must be justified or explained, i.e., relative to the number of features that the conclusion category is known to possess.

Definition of Similarity

As mentioned above, the feature-based model does not assume that similarity is computed in the course of making argument strength judgments. Nevertheless, in order to make contact with the phenomena to be described, a theoretical measure of the similarity between two category vectors is required. This model of similarity does not reflect any of the

computations performed by the feature-based model. It functions only to represent phenomena described in terms of similarity in a language compatible with the feature-based model.

Perhaps the simplest measure of similarity would be the dot product between the two category vectors, an indicator of the categories' common features. One disadvantage of the dot product is its sensitivity to the magnitudes of its operands. Consequently, I define the *similarity* (sim) of two categories D and E using a measure of correlation between the vectors $F(D)$ and $F(E)$, specifically the cosine of the angle between them:

$$\text{sim}(D,E) = \frac{F(D) \cdot F(E)}{|F(D)|\,|F(E)|}. \tag{6}$$

The virtues of this measure are that it captures most of the properties normally attributed to similarity judgments and its relation to the feature-based model is simple mathematically. This will facilitate many of the demonstrations below concerning phenomena involving similarity judgments.

As a model of similarity, Equation 6 shares with Tversky's (1977) contrast model the assumption that similarity is proportional to two categories' common features and inversely proportional to their distinctive features. I compared the two models by correlating their predictions of the similarities between 1128 mammal pairs using the Osherson et al. (1991) feature ratings. The correlation was 0.97 between the values given by Eq. 6 and those of an additive version of the contrast model which gave equal weight to common and distinctive features. Clearly the models are closely related. Equation 6 differs from the general contrast model in assuming similarity to be a symmetric relation and in assigning self-similarity, the similarity between a category and itself, the maximum value of 1. Neither of these assumptions will play a role in the explanations I provide below for any of the phenomena. I am not claiming that Eq. 6 is the best model of similarity available. The intention of Eq. 6 is only to describe a relation which, when it holds, has certain consequences for the feature-based model. It serves as a link between the model and phenomena described in terms of similarity.

Since $F(D)$ and $F(E)$ may depend on the context in which they are presented, $\text{sim}(D,E)$ may vary. Because I assume that category representations are constant within each of the experiments to be discussed, similarities should not vary either. Notice that, unlike Osherson et al.'s (1990) SIM_S function, the similarity between *any* two categories is well-defined, even categories usually thought to reside at different hierarchical levels.

PHENOMENA CONSISTENT WITH BOTH MODELS

Osherson et al. (1990) and Smith et al. (1992) review, propose, and

provide empirical support for several phenomena concerning the variables that influence argument strength. Since the phenomena are intended to describe tendencies, as opposed to laws, we should expect counterexamples to exist, as both models predict. In fact, the phenomena themselves can generate opposing expectations for certain arguments (e.g., phenomenon *i* below, similarity, versus phenomena *ii* and *iii*, premise diversity). Because of this fuzziness in the sets of arguments to which they apply, the phenomena are presented by example, and I demonstrate how the feature-based model accounts for each of these examples. Each phenomenon is illustrated with one or more contrasting pairs of arguments in which the first argument is stronger than the second.

Osherson et al. (1990) distinguish among general, specific, and mixed arguments. Arguments are *general* if premise categories are all properly included in the conclusion category, for example

Mice have *X*.
Antelopes have *X*.

Mammals have *X*.

Arguments are *specific* if any category that properly includes one of the premise or conclusion categories also properly incudes the others, for example

Pigs have *X*.
Antelopes have *X*.

Aardvarks have *X*.

Here, Mammal properly includes Pigs, Antelopes, and Aardvarks. An argument is called *mixed* if it is not general or specific, for example

Mice have *X*.
Flamingoes have *X*.

Mammals have *X*.

Each of these argument types raises slightly different concerns for the similarity-coverage model. This led Osherson et al. (1990) to distinguish phenomena that differ only in the type of argument to which they apply. Because the feature-based model does not distinguish the representation of a general category from that of a specific category in that all categories are represented as vectors of features, many of these concerns do not arise. I therefore will limit my demonstrations to only one of the argument types, but maintain the counting scheme employed by Osherson et al. to number phenomena in order to facilitate comparison between the models.

i. Premise–Conclusion Similarity (Specific)

Arguments are strong to the extent that categories in the premises are similar to the conclusion category (Rips, 1975). The following exemplifies the phenomenon for single-premise arguments:

A. German shepherds have *X*. / Collies have *X*. is stronger than B. German shepherds have *X*. / Chihuahuas have *X*.

Sixty-one of 64 University of Michigan undergraduates chose the first argument as more convincing ($p < .0001$).[2] Similarity ratings were also collected from 63 of these subjects. The mean similarity rating between German shepherds and Collies was 5.30, significantly higher than that between German shepherds and Chihuahuas (3.88), $t(62) = 7.38$, $p < .0001$. The feature-based model attributes this phenomenon to feature coverage: Due to their greater similarity, German shepherds cover more features of Collies than they do of Chihuahuas. A more formal analysis follows.

To obtain argument strengths, the first step is to encode the premise German shepherds (GS) have *X*. From (4), W(GS) = F(GS). The second step is to test the conclusion. From Eq. 5, the strength of Argument A is

$$a_x(\text{Collies/GS}) = \frac{\text{F(GS)} \cdot \text{F(Collies)}}{|\text{F(Collies)}|^2}.$$

The strength of Argument B is derived similarly with the result that

$$a_x(\text{Chihuahuas/GS}) = \frac{\text{F(GS)} \cdot \text{F(Chihuahuas)}}{|\text{F(Chihuahuas)}|^2}.$$

One way that we can determine whether the feature-based model makes the right prediction is to compute argument strengths from empirically determined estimates of category features and then compare these strengths to the obtained ones. I used Osherson et al.'s (1991) ratings to

[2] The blank predicate used for both this argument and the multiple-premise example was "produce THS by their pituitary." Unless noted otherwise, the data reported were gathered from students who were given questionnaires which asked them to either choose between pairs of arguments or rate single arguments according to their convincingness. A pair of arguments bearing no relation to any tested argument was provided as an example and briefly discussed to clarify the meaning of "convincingness." Subjects were asked to consider each argument on its own merit, ignoring all information other than the facts presented. Following the argument strength task, similarity ratings were collected from the same students on a 7-point scale where 1 was "not at all" and 7 was "very" similar. Students went through the questionnaire at their own pace and were given as much time as they desired.

estimate F(GS), F(Collies), and F(Chihuahuas), and computed that F(GS) · F(Collies) = 7.5 and F(GS) · F(Chihuahuas) = 6.8; the former pair of categories have more common features. Furthermore, $|F(\text{Collies})|^2 = 7.7$, slightly less than $|F(\text{Chihuahuas})|^2 = 8.0$. Mainly because German shepherds and Collies have more common features than do German shepherds and Chihuahuas, the feature-based model correctly predicts that $a_x(\text{Collies/GS}) > a_x(\text{Chihuahuas/GS})$.

A more general formulation of the model's predictions can be obtained by examining its relation to Eq. 6, our model of similarity. Consider categories A, B, and C such that $\text{sim}(A,B) > \text{sim}(A,C)$ or, from Eq. 6,

$$\frac{F(A) \cdot F(B)}{|F(A)|\,|F(B)|} > \frac{F(A) \cdot F(C)}{|F(A)|\,|F(C)|},$$

which implies

$$\frac{F(A) \cdot F(B)}{|F(B)|} > \frac{F(A) \cdot F(C)}{|F(C)|}. \tag{7}$$

What are the implications of this inequality for the relative strength of the arguments A have X, therefore B have X and A have X, therefore C have X? The strength of the former argument is

$$\begin{aligned} a_x(B/A) &= \frac{F(A) \cdot F(B)}{|F(B)|^2} && \text{by Eq. 5,} \\ &> \frac{F(A) \cdot F(C)}{|F(C)|\,|F(B)|} && \text{by substitution with Eq. 7,} \\ &\geq \frac{F(A) \cdot F(C)}{|F(C)|^2} && \text{if } |F(C)| \geq |F(B)| \\ &= a_x(C/A) && \text{by Eq. 5.} \end{aligned}$$

$a_x(C/A)$ is the strength of the second argument. So as long as $|F(C)| \geq |F(B)|$, the argument involving the more similar pair of categories will be stronger. In most cases, this argument will be stronger even if $|F(B)| > |F(C)|$. To see this, observe that rearranging Eq. 7 gives

$$F(A) \cdot F(B) > \frac{|F(B)|}{|F(C)|} F(A) \cdot F(C).$$

In words, A and B have more common features than A and C relative to the ratio of the magnitudes of B and C. If A and B have sufficiently many more common features than A and C, if $F(A) \cdot F(B)$ is sufficiently greater

than $F(A) \cdot F(C)$, then the right-hand side of this inequality will remain smaller even if we square the coefficient $|F(B)|/|F(C)|$. In this case, we have that

$$F(A) \cdot F(B) > \left[\frac{|F(B)|}{|F(C)|}\right]^2 F(A) \cdot F(C).$$

If this condition does hold, then by dividing both sides by $|F(B)|^2$, we have that $a_x(B/A) > a_x(C/A)$, or the first argument is stronger, even if $|F(B)| > |F(C)|$. The critical condition for the similarity phenomenon to hold is that the features of the more similar conclusion category are better covered by the premise category. If the features of the two more similar categories have enough overlap, they will produce a stronger argument regardless of the magnitude of the conclusion category.

As an illustration of the power of this analysis, I now provide a counterexample to the similarity phenomenon that is consistent with both the feature-based model and Eq. 6, the model of similarity. Consider the pair of arguments

$$\frac{\text{Fact: Bobcats have sesamoid bones.}}{\text{Conclusion: Moles have sesamoid bones.}} \quad \text{(e)}$$

$$\frac{\text{Fact: Bobcats have sesamoid bones.}}{\text{Conclusion: Rats have sesamoid bones.}} \quad \text{(f)}$$

According to the judgments of 40 University of Michigan students, the similarity of Bobcats and Moles, which averaged 3.10, was not significantly different from the mean similarity rating of Bobcats and Rats (2.95), $t(39) = 1.03$, ns. Equation 6 expresses this fact as

$$\frac{F(\text{Bobcats}) \cdot F(\text{Moles})}{|F(\text{Bobcats})|\,|F(\text{Moles})|} = \frac{F(\text{Bobcats}) \cdot F(\text{Rats})}{|F(\text{Bobcats})|\,|F(\text{Rats})|},$$

which implies

$$\frac{F(\text{Bobcats}) \cdot F(\text{Moles})}{|F(\text{Moles})|} = \frac{F(\text{Bobcats}) \cdot F(\text{Rats})}{|F(\text{Rats})|}.$$

According to Osherson et al.'s (1991) feature ratings, more properties were rated as having a higher strength of association to Rats than to Moles, $|F(\text{Rats})| > |F(\text{Moles})|$, which is consistent with the claim that people's representation of Rats is, on average, richer than that of Moles. In conjunction with the equality immediately above, this implies

$$\frac{F(\text{Bobcats}) \cdot F(\text{Moles})}{|F(\text{Moles})|^2} > \frac{F(\text{Bobcats}) \cdot F(\text{Rats})}{|F(\text{Rats})|^2},$$

or that, according to the feature-based model, Argument e should be stronger than Argument f and 26 of 40 of the students chose it the stronger ($p < .05$ by a binomial test of the probability of achieving a value of 26 or greater).[3] This pattern of data was replicated with a separate group of subjects. In short, the feature-based model was able to predict a case in which one argument was stronger than another, despite apparently equal category similarity, by virtue of the relative magnitudes of their conclusions.

The problem is more complicated with two-premise argument pairs such as

$$\frac{\text{A. German shepherds have } X. \quad \text{Dalmatians have } X.}{\text{Collies have } X.} \text{ is stronger than } \frac{\text{B. German shepherds have } X. \quad \text{Dalmatians have } X.}{\text{Chihuahuas have } X.},$$

in which both German shepherds and Dalmatians are more similar to Collies than to Chihuahuas (60 of 64 students chose Argument A as the stronger one. The mean similarity of Dalmatians and Collies was judged to be 4.94, significantly greater than that of Dalmatians and Chihuahuas of 3.86, $t(63) = 6.08$, $p < .0001$). The explanation for the premise–conclusion similarity phenomenon with multiple premises is conceptually identical to that for the single-premise case. Argument strength is proportional to the dot product of the conclusion category vector and the weight vector encoding the premises. The more features shared by the conclusion and premise categories, the greater will be this dot product. Let GS stand for German shepherds, D for Dalmatians, C for Collies, and Ch for Chihuahuas. In Appendix B, I show that we can write the strength of Argument A as

$$a_x(\text{C/GS,D}) = \frac{\text{F(GS)} \cdot \text{F(C)} + [1 - a_x(\text{D/GS})][\text{F(D)} \cdot \text{F(C)} - \Sigma f_i(\text{GS}) f_i(\text{D}) f_i(\text{C})]}{|\text{F(C)}|^2}$$

$$= \frac{\text{F(GS)} \cdot \text{F(C)}}{|\text{F(C)}|^2} + [1 - a_x(\text{D/GS})] \frac{\text{F(D)} \cdot \text{F(C)}}{|\text{F(C)}|^2} - [1 - a_x(\text{D/GS})] \frac{\Sigma f_i(\text{GS}) f_i(\text{D}) f_i(\text{C})}{|\text{F(C)}|^2}.$$

The strength of Argument B, derived in identical manner, is

[3] All p values reported below corresponding to choice proportions will refer to binomial tests.

$$a_x(\text{Ch/GS,D}) = \frac{\text{F(GS)} \cdot \text{F(Ch)}}{|\text{F(Ch)}|^2} + [1 - a_x(\text{D/GS})] \frac{\text{F(D)} \cdot \text{F(Ch)}}{|\text{F(Ch)}|^2} - [1 - a_x(\text{D/GS})] \frac{\Sigma f_i(\text{GS}) f_i(\text{D}) f_i(\text{Ch})}{|\text{F(Ch)}|^2}.$$

Argument strength increases with the featural overlap of both premise categories to the conclusion category as projected in their dot products (e.g., F(GS) · F(C) and F(D) · F(C) in Argument A). It decreases with the extent to which premise and conclusion categories all share features as projected in the three-term products (e.g., $\Sigma f_i(\text{GS}) f_i(\text{D}) f_i(\text{C})$). This holds because a premise only increases belief in a conclusion to the extent that the features its category shares with the conclusion category are not redundant; i.e., they are not also shared by other premise categories. Because feature values are all restricted to be less than or equal to one, an argument's three-term product is necessarily less than or equal to that argument's dot product terms. Because of the large number of dimensions assumed to be involved, and because features are not necessarily binary, only in very unusual cases would equality hold.

According to Osherson et al.'s (1991) feature ratings, the required conditions are satisfied for this example and the feature-based model predicts that Argument A is stronger than B.[4] To see why we expect the similarity phenomenon to generally hold for two-premise arguments, let GS, D, C, and Ch represent generic categories in which GS and C are more similar than GS and Ch, while D and C are more similar than D and Ch. Then, according to the reasoning for single-premise arguments, the two dot product terms in Argument A will normally be greater than the corresponding terms in Argument B. Even if the three-term product is smaller in Argument B than Argument A, the difference is unlikely to be as great as the combined difference in the dot products. Moreover, unlike the first dot product term, the effect of the three-way product is diminished by the coefficient preceding it $[1 - a_x(\text{D/GS})]$. Therefore, as long as the magnitude of the C vector is not too much greater than that of the Ch vector, Argument A is likely to be stronger than Argument B.

Analytic derivations are too complicated to detail for three or more premise arguments. Their patterns, however, show a clear generalization of the one- and two-premise cases and the same intuitions hold. Although this increase in derivational complexity with the number of premises may be considered a drawback of the feature-based model, note that the computational complexity of the model is only linearly related to the number

[4] $\text{F(GS)} \cdot \text{F(C)} = 7.5 > \text{F(GS)} \cdot \text{F(Ch)} = 6.8$, $\text{F(D)} \cdot \text{F(C)} = 6.2 > \text{F(D)} \cdot \text{F(Ch)} = 5.6$, $\Sigma f_i(\text{GS}) f_i(\text{D}) f_i(\text{C}) = 3.3 > \Sigma f_i(\text{GS}) f_i(\text{D}) f_i(\text{Ch}) = 3.1$. $|\text{F(C)}|^2 = 7.7 < |\text{F(Ch)}|^2 = 8.5$.

of premises. The same encoding rule, Eq. 1, is applied iteratively to each premise. Perhaps more important, the conceptual complexity of the model is unrelated to the number of premises. The same rules are applied for the same reasons to any argument, whether it has one or many premises. Arguments are deemed strong in all cases if the weights obtained after encoding the premises cover the features of the conclusion category. Although I will not discuss them analytically, the model's predictions for the three-premise case show good fits to data, as discussed in the section on quantitative tests below.

ii and iii. Premise Diversity (General and Specific)

The less similar premises are to each other, the stronger the argument tends to be. An example of a pair of general arguments satisfying premise diversity is

A.		B.
Hippos have *X*.		Hippos have *X*.
Hamsters have *X*.	is stronger than	Rhinos have *X*.
Mammals have *X*.		Mammals have *X*.

For an example of a pair of specific arguments, substitute Giraffes for Mammals (cf. Osherson et al., 1990). The account of the specific case follows from that of the general one. The feature-based model attributes the diversity phenomenon, again, to feature coverage. More diverse premises cover the space of features better than more similar premises because their features are not redundant, and therefore are more likely to overlap with features of the conclusion category and more likely to be encoded. More precisely, (i) featural overlap between dissimilar premise categories and a conclusion category will tend to be less redundant than the overlap between similar premise categories and a conclusion category; and (ii) weight changes are greater during encoding of a dissimilar premise than a similar premise (dissimilar premises have more surprise value). A more formal analysis follows.

The similarity of Hippos and Rhinos is greater than that of Hippos and Hamsters. In the discussion of Phenomenon *i*, similarity, I showed that this leads us to expect

$$a_x(\text{Rhinos/Hippos}) > a_x(\text{Hamsters/Hippos}). \tag{8}$$

Let Hi stand for Hippos, Ha for Hamsters, M for Mammals, and R for Rhinos. As shown in Appendix B, we can write the strength of Argument A as

$$a_x(\text{M/Hi,Ha}) = \frac{\text{F(Hi)} \cdot \text{F(M)} + [1 - a_x(\text{Ha/Hi})][\text{F(Ha)} \cdot \text{F(M)} - \Sigma f_i(\text{Hi}) f_i(\text{Ha}) f_i(\text{M})]}{|\text{F(M)}|^2}$$

and the strength of Argument B as

$$a_x(\text{M/Hi,R}) = \frac{\text{F(Hi)} \cdot \text{F(M)} + [1 - a_x(\text{R/Hi})][\text{F(R)} \cdot \text{F(M)} - \Sigma f_i(\text{Hi}) f_i(\text{R}) f_i(\text{M})]}{|\text{F(M)}|^2}.$$

Cancelling like terms, we see that Argument A is stronger than Argument B if and only if

$$[1 - a_x(\text{Ha/Hi})][\text{F(Ha)} \cdot \text{F(M)} - \Sigma f_i(\text{Hi}) f_i(\text{Ha}) f_i(\text{M})] > [1 - a_x(\text{R/Hi})][\text{F(R)} \cdot \text{F(M)} - \Sigma f_i(\text{Hi}) f_i(\text{R}) f_i(\text{M})]. \quad (9)$$

Inequality 8 states that $[1 - a_x(\text{Ha/Hi})] > [1 - a_x(\text{R/Hi})]$, so the first term on the left hand side of Condition 9 is greater than the first term on the right. $\text{F(R)} \cdot \text{F(M)}$ is a measure of the features that Rhinos have in common with the mental representation—the prototype perhaps—of the Mammal category. Most of us comparing these arguments do not know substantially more about Rhinos than about Hamsters, which is consistent with Osherson et al.'s feature ratings, $|\text{F(R)}| = 2.6 < |\text{F(H)}| = 3.0$. Moreover, Rhinos are not more typical Mammals than are Hamsters, as evidenced by both the rarity of horned mammals with plated skin, and typicality judgments showing that in fact Hamsters are more typical.[5] We can expect, therefore, that for most of us $\text{F(R)} \cdot \text{F(M)}$ is not substantially greater, and probably less, than $\text{F(Ha)} \cdot \text{F(M)}$. The only remaining difference between the two sides of the inequality lies in the three-way product terms $\Sigma f_i(\text{Hi}) f_i(\text{Ha}) f_i(\text{M})$ and $\Sigma f_i(\text{Hi}) f_i(\text{R}) f_i(\text{M})$. The value of these terms would tend to be proportional to premise similarity because the more features two-premise categories share, the more likely that features will be shared by all of the premise and conclusion categories. Because the premises of Argument B are more similar, $\Sigma f_i(\text{Hi}) f_i(\text{R}) f_i(\text{M})$ is likely to be greater than $\Sigma f_i(\text{Hi}) f_i(\text{Ha}) f_i(\text{M})$. So a bigger quantity is being subtracted on the right-hand side and Condition 9 is therefore likely to hold. More precisely, Argument A will be stronger than Argument B for every subject for whom it does hold.

Feature Exclusion (New Phenomenon)

The feature-based model predicts a boundary condition on the diversity phenomenon. A premise category that has no overlap with the conclusion

[5] Unpublished data collected by Tony Wilkie.

category should have no effect on argument strength even if it leads to a more diverse set of premises. Consider the arguments

Foxes require trace amounts of magnesium for reproduction.
Deer require trace amounts of magnesium for reproduction.

Weasels require trace amounts of magnesium for reproduction. (g)

and

Foxes require trace amounts of magnesium for reproduction.
Rhinos require trace amounts of magnesium for reproduction.

Weasels require trace amounts of magnesium for reproduction. (h)

The second argument indeed has a more diverse set of premises than the first; the similarity between Foxes and Rhinos was rated significantly lower than the similarity between Foxes and Deer (2.02 and 4.00, respectively, $t(45) = 10.00$; $p < .001$). Nevertheless, the feature-based model predicts that Argument g will be stronger than Argument h because Rhinos and Weasels have so few features in common, the dot product of their feature vectors was only 3.0 in the feature ratings collected by Osherson et al. (1991), relative to Deer and Weasels (dot product of 5.5). Let F be Foxes, R be Rhinos, and W be Weasels. The derivation in Appendix B tells us that the argument strength of h is

$$a_x(\mathrm{W/F,R}) = \frac{\mathrm{F(F)} \cdot \mathrm{F(W)} + [1 - a_x(\mathrm{R/F})][\mathrm{F(R)} \cdot \mathrm{F(W)} - \Sigma f_i(\mathrm{F}) f_i(\mathrm{R}) f_i(\mathrm{W})]}{|\mathrm{F(W)}|^2}.$$

If $\mathrm{F(R)} \cdot \mathrm{F(W)}$ is negligible, then so is $\Sigma f_i(\mathrm{F}) f_i(\mathrm{R}) f_i(\mathrm{W})$ since it is always positive but less than $\mathrm{F(R)} \cdot \mathrm{F(W)}$ which implies that

$$\begin{aligned} a_x(\mathrm{W/F,R}) &\approx \frac{\mathrm{F(F)} \cdot \mathrm{F(W)}}{|\mathrm{F(W)}|^2} \\ &= a_x(\mathrm{W/F}). \end{aligned}$$

In other words, Rhinos contribute little strength to the argument, even though Foxes and Rhinos compose a diverse set. To test this analysis, I asked 46 University of Michigan undergraduates to choose the stronger of Arguments g and h. As predicted, 41 of them chose g ($p < .0001$).

The similarity–coverage model can explain this result by assuming that, even though the premises in h had greater diversity, they had less coverage. If Rhinos have few similar neighbors, they may not add much to

the coverage of the lowest-level category that includes both premises and conclusion.[6] To derive the prediction from the similarity–coverage model, we need first to define the lowest-level category both with and without Rhinos as a premise (presumably Mammals in both cases) and then compute coverage using similarity ratings between every member of this lowest-level category and each premise category. To derive the prediction from the feature-based model, we need to know that the feature overlap between the premise and conclusion categories is small.

I have replicated the above demonstration using different pairs of arguments. Here is one more example:

Fact: Foxes secrete uric acid crystals.
Fact: Beavers secrete uric acid crystals.
Conclusion: Chihuahuas secrete uric acid crystals. (i)

Fact: Foxes secrete uric acid crystals.
Fact: Humpback whales secrete uric acid crystals.
Conclusion: Chihuahuas secrete uric acid crystals. (j)

The diversity between the premises of Argument j is greater than that in Argument i; the rated similarity between Foxes and Humpback whales was 2.24, significantly lower than that between Foxes and Beavers (4.16), $t(57) = 10.97$, $p < .001$. Nevertheless, 56 of 59 students chose i. The feature-based model predicted this on the basis of the observation that the dot product between the feature vector associated with Humpback whales and that associated with Chihuahuas was only 2.0.

iv. Premise Typicality (General)

The more typical premise categories are of the conclusion category, the stronger is the argument (Rothbart & Lewis, 1988). For example,

A. Wolves have *X*. / Mammals have *X*. is stronger than B. Oxen have *X*. / Mammals have *X*.,

[6] Supporters of the similarity–coverage model cannot appeal to greater similarity between the premises and conclusion of Argument g relative to Argument h to explain this result. According to that model, overall similarity is given by the maximum of the pairwise similarities between each premise category and the conclusion category. Yet, only 1 of 46 subjects rated Deer more similar to Weasels than Foxes to Weasels. With regard to the following demonstration, only 7 of 39 subjects rated Beavers more similar to Chihuahuas than Foxes to Chihuahuas. Excluding these subjects from the analysis has no effect on the pattern of data.

where Wolves are more typical of the category Mammals than are Oxen.[7] By virtue of the typicality of Wolves, their similarity to Mammals is greater than the similarity of Oxen to Mammals. Therefore, from Eq. 6,

$$\begin{aligned}\text{sim(Wolves,Mammals)} &= \frac{\text{F(Wolves)} \cdot \text{F(Mammals)}}{|\text{F(Wolves)}|\,|\text{F(Mammals)}|} \\ &> \frac{\text{F(Oxen)} \cdot \text{F(Mammals)}}{|\text{F(Oxen)}|\,|\text{F(Mammals)}|} \\ &= \text{sim(Oxen,Mammals)}.\end{aligned}$$

Tversky (1977) proposed that the features of more typical items are more salient or prominent. The typicality phenomenon can be derived directly from the difference in similarity for those pairs of categories that satisfy a weaker condition, namely that the representation of the more typical category is at least as rich as that of the more atypical category. This condition will hold for most readers for this example because Wolves tend to be more familiar than Oxen. Osherson et al.'s (1991) feature ratings also provide support, $|\text{F(Wolves)}| = 3.4 > |\text{F(Oxen)}| = 3.2$. This condition, along with the inequality above, implies that

$$\frac{\text{F(Wolves)} \cdot \text{F(Mammals)}}{|\text{F(Mammals)}|^2} > \frac{\text{F(Oxen)} \cdot \text{F(Mammals)}}{|\text{F(Mammals)}|^2}. \qquad (10)$$

Notice that Eq. 10 will hold as long as Wolves have more features in common with Mammals than Oxen do, regardless of their respective magnitudes. Taken generally, condition 10 will obtain whenever a typical category has more in common with its superordinate than does an atypical category. To derive the strength of Argument A, we use Eq. 5:

$$a_x\text{(Mammals/Wolves)} = \frac{\text{F(Wolves)} \cdot \text{F(Mammals)}}{|\text{F(Mammals)}|^2}.$$

The strength of Argument B is

$$a_x\text{(Mammals/Oxen)} = \frac{\text{F(Oxen)} \cdot \text{F(Mammals)}}{|\text{F(Mammals)}|^2}.$$

The inequality in Eq. 10 dictates that a_x(Mammals/Wolves) > a_x(Mammals/Oxen) or that Argument A is stronger than Argument B.

As an illustration of this analysis, I provide an example in which argu-

[7] The blank predicate was "use serotonin as a neurotransmitter." Twenty-nine of 39 students chose the first argument ($p < .01$). The mean typicality judgment for Wolves as Mammals was 5.23, for Oxen it was only 4.85, $t(38) = 2.84$, $p < .01$.

ment strength is derived using categories which are equally similar to a superordinate, but which differ in their magnitudes. Consider the arguments

$$\frac{\text{Fact: Bobcats have potassium in their cerebral fluid.}}{\text{Conclusion: Mammals have potassium in their cerebral fluid.}} \quad \text{(k)}$$

and

$$\frac{\text{Fact: Weasels have potassium in their cerebral fluid.}}{\text{Conclusion: Mammals have potassium in their cerebral fluid.}} \quad \text{(l)}$$

Bobcats and Weasels were judged to be equally similar to Mammals (4.74 and 4.87, respectively; $t < 1$). According to Eq. 6,

$$\begin{aligned}\text{sim(Bobcats,Mammals)} &= \frac{\text{F(Bobcats)} \cdot \text{F(Mammals)}}{|\text{F(Bobcats)}|\,|\text{F(Mammals)}|} \\ &= \frac{\text{F(Weasels)} \cdot \text{F(Mammals)}}{|\text{F(Weasels)}|\,|\text{F(Mammals)}|} \\ &= \text{sim(Weasels,Mammals).}\end{aligned}$$

However, Osherson et al.'s (1991) feature ratings indicate that Bobcats have a richer representation than Weasels, $|\text{F(Bobcats)}| = 3.4 > |\text{F(Weasels)}| = 2.5$, which implies, in conjunction with the equality in similarities, that Bobcats and Mammals have more features in common than Weasels and Mammals or that

$$\frac{\text{F(Bobcats)} \cdot \text{F(Mammals)}}{|\text{F(Mammals)}|^2} > \frac{\text{F(Weasels)} \cdot \text{F(Mammals)}}{|\text{F(Mammals)}|^2},$$

or a_x(Mammals/Bobcats) > a_x(Mammals/Weasels). Argument k should be stronger than argument l and 32 of 39 subjects judged it to be ($p < .0001$). The feature-based model successfully predicted how comparable similarity judgments can combine with unequal magnitude estimates to make one argument stronger than another.

v and vi. Premise Monotonicity (General and Specific)

Adding a premise whose category is new and chosen from the lowest level category that includes both the categories of the old premises and the conclusion will increase the strength of an argument. A pair of general arguments that satisfy premise monotonicity is

$$\frac{\begin{array}{l}\text{A.}\\ \text{Sparrows have } X.\\ \text{Eagles have } X.\\ \text{Hawks have } x.\end{array}}{\text{Birds have } X.} \quad \text{is stronger than} \quad \frac{\begin{array}{l}\text{B.}\\ \text{Sparrows have } X.\\ \text{Eagles have } X.\end{array}}{\text{Birds have } X.}.$$

An example of a pair of specific arguments satisfying premise monotonicity would be obtained by substituting "Ravens" for "Birds" in both arguments. The analysis below would apply equally well. Premise monotonicity holds whenever the conclusion category shares values on one or more features with the additional premise category that it does not share with the other premise categories. Specifically, the new premise category must include features of the conclusion category whose corresponding weights are not already at their maximum value after encoding the old premise categories.

More formally, the strength of Argument A is

$$a_x(\text{Birds/Hawks,Sparrows,Eagles}) = \frac{\text{W(Hawks,Sparrows,Eagles)} \cdot \text{F(Birds)}}{|\text{F(Birds)}|^2},$$

and the strength of Argument B is

$$a_x(\text{Birds/Sparrows,Eagles}) = \frac{\text{W(Sparrows,Eagles)} \cdot \text{F(Birds)}}{|\text{F(Birds)}|^2}.$$

Recall that weights are never decreased (assuming $a_x \leq 1$). Adding a new premise will always increase weights connected to features of its category that are not already at their maximum value. Therefore, for each i,

$$w_i(\text{Hawks,Sparrows,Eagles}) \geq w_i(\text{Sparrows,Eagles}),$$

which implies that

$$\frac{\text{W(Hawks,Sparrows,Eagles)} \cdot \text{F(Birds)}}{|\text{F(Birds)}|^2} \geq \frac{\text{W(Sparrows,Eagles)} \cdot \text{F(Birds)}}{|\text{F(Birds)}|^2}.$$

In terms of a_x, or argument strength,

$$a_x(\text{Birds/Hawks,Sparrows,Eagles}) \geq a_x(\text{Birds/Sparrows,Eagles})$$

and strictly greater whenever

$$[\text{W(Hawks,Sparrows,Eagles)} - \text{W(Sparrows,Eagles)}] \cdot \text{F(Birds)} > 0.$$

Although the contribution to argument strength provided by the new premise may be small, it is necessarily nonnegative. The model therefore predicts that when the monotonicity phenomenon is evaluated over a large number of subjects, some tendency to prefer the argument with the larger number of premises should be observed.

vii. Inclusion Fallacy (One Specific and One General Argument)

Shafir, Smith, and Osherson (1990) demonstrate that, counter to normative prescription, arguments with more general conclusions can sometimes seem stronger than arguments with identical premises but specific conclusions:

$$\frac{\text{Robins have } X.}{\text{Birds have } X.} \text{ is stronger than } \frac{\text{Robins have } X.}{\text{Ostriches have } X.}$$

(A. B.)

The feature-based model explains this phenomenon by appealing to greater coverage by the premise category of the features of the general category (Birds) than of the specific category (Ostriches). The account follows directly from that of phenomenon *i*, similarity, on the assumption that Robins are more similar to Birds than they are to Ostriches. This is precisely the assumption made by Osherson et al. (1990). The feature-based model makes no distinction between the representation of general and specific categories; both are represented as feature vectors. Therefore, any of the model's logic that applies to one kind of category will apply equally to the other.

viii. Conclusion Specificity (General)

When premise categories are properly included in the conclusion category, arguments tend to be stronger the more specific is the conclusion category. Corroborating data can be found in Rothbart and Lewis (1988). This is illustrated by the following example because Birds is more specific than Animals, and because Sparrows and Eagles are Birds which are in turn Animals

$$\frac{\text{Sparrows have } X. \quad \text{Eagles have } X.}{\text{Birds have } X.} \text{ is stronger than } \frac{\text{Sparrows have } X. \quad \text{Eagles have } X.}{\text{Animals have } X.}$$

(A. B.)

This phenomenon follows from *i*, premise–conclusion similarity, whenever the more specific category is more similar than the more general category to the premise categories. Instances of Birds tend to be more

similar to the category of Birds than to the more general category of Animals (cf. Smith, Shoben, & Rips, 1974), implying that Birds are more similar to Sparrows and Eagles than Animals are so that Argument A will be stronger than Argument B.

However, categories k levels apart are not always more similar than categories more than k levels apart (Smith et al., 1974). If a premise category is more similar to a more general category, the feature-based model expects the conclusion specificity phenomenon to reverse as long as the magnitude of the vector corresponding to the more general category is not too large. For instance, Chickens are more similar to Animals than to Birds for some people. The model predicts that the Argument "Chickens have X, therefore Animals have X" will be stronger for these people than "Chickens have X, therefore Birds have X." Evidence for just this type of event is provided by the inclusion fallacy. Gelman (1988) has offered an alternative explanation for conclusion specificity which appeals to the relative homogeneity of categories at different levels.

ix. Premise–Conclusion Asymmetry (Specific)

This phenomenon, first discussed by Rips (1975), involves single-premise arguments. Switching premise and conclusion categories can lead to asymmetric arguments, in the sense that the strength of P/C will differ from that of C/P. An example is provided by Osherson et al. (1990):

$$\frac{\text{A. Mice have } X.}{\text{Bats have } X.} \text{ is stronger than } \frac{\text{B. Bats have } X.}{\text{Mice have } X.}.$$

Those authors, along with Rips (1975), attribute the phenomenon to differences in typicality. People are assumed to be more willing to generalize from a typical instance to an atypical instance than from an atypical instance to a typical instance. According to the similarity–coverage model, this is because typical instances provide greater coverage of the lowest-level category that includes both the premise and conclusion categories. The phenomenon follows if Mice are assumed more typical than Bats of some common category, such as Mammals.

The feature-based model attributes the phenomenon to differential richness of category representations (which is likely to be correlated with typicality). From Eq. 5, we write the strength of Argument A as

$$a_x(\text{Bats/Mice}) = \frac{\text{F(Mice)} \cdot \text{F(Bats)}}{|\text{F(Bats)}|^2},$$

and the strength of Argument B as

$$a_x(\text{Mice/Bats}) = \frac{\text{F(Bats)} \cdot \text{F(Mice)}}{|\text{F(Mice)}|^2}.$$

Because their numerators are identical, the relative strength of the two arguments depends entirely on the relative magnitudes of their conclusion categories. On the one hand, Bats have distinctive properties that would tend to increase the richness of their representations, like having wings, being nocturnal, and living in caves. On the other hand, we know a lot about Mice because they are so familiar (e.g., they have long tails, they eat cheese, we have a more precise idea of their appearance) and this gives us confidence that they possess common mammalian properties that we know about (e.g., they probably have normal sensory and motor systems). In any case, the relative magnitudes of the two categories' representations are an open (and difficult) empirical question, the complexity of which cannot be addressed with simple feature ratings.

However, categories do exist which will let us both make reasonable guesses as to the relative magnitude of their representations and also pit the two explanations for the asymmetry phenomenon against each other. Given a pair of categories D and E in which the representation of D is richer than that of E but E is more typical than D of the lowest-level category that properly includes them both, the argument D/E should be stronger than E/D. The similarity–coverage model predicts the opposite.

I chose six such pairs of categories (Killer whales and Weasels, Kangaroos and Wolverines, Chickens and Orioles, Penguins and Finches, Tomatoes and Papayas, and Cucumbers and Guavas). In each case, the first category is less typical of the lowest-level inclusive category than the second (Mammals, Mammals, Birds, Birds, Fruit, and Fruit, respectively), but it is more familiar and subjects are likely to have more detailed knowledge of it.[8] I constructed two arguments from each pair of categories by having each category serve as both premise and conclusion category and asked subjects to rate the convincingness of each argument on a scale from 0 to 10 (see Appendix A for the design and procedure).

For five of the six pairs, the argument with the less typical premise and lower magnitude conclusion category (e.g., Chickens have *X*, therefore Orioles have *X*) was judged more convincing (mean of 3.12) than the argument with the more typical premise and greater magnitude conclusion category (e.g., Orioles have *X*, therefore Chickens have *X*; mean of 2.93). The sole exception was the pair involving Killer whales and Weasels. Although the judgments were in a direction that supported the feature-

[8] The Wolverine is the University of Michigan mascot and therefore could be more familiar to our subjects than the Kangaroo. However, informal questioning of the participants suggested otherwise.

based model, the overall difference was not statistically significant for these item–pairs across subjects, $t(33) = 1.54$, ns. The experiment may not have had enough power to detect a real difference between the two types of arguments. Using a larger number of arguments with familiar but nonexplainable predicates, Sloman and Wisniewski (1992) have found an asymmetry effect in which subjects preferred arguments with a lower magnitude conclusion category. Another possibility is that both representational richness and typicality influence argument strength and that they had opposite effects in this experiment. Typicality may influence performance to the extent that subjects' reasoning is logical and not intuitive. They may employ an inferential step of the form "If the property is true of such a representative member of the category, then it is likely to be true of all category members," an inference that is nicely captured by the similarity–coverage model.

x. Premise–Conclusion Identity

Phenomena *x* and *xi* were dubbed "limiting-case phenomena" and posited by Osherson et al. (1990, 1991) without experimental evidence. Premise–conclusion identity states that arguments with identical premises and conclusions are perfectly strong, i.e., they have maximum argument strength:

$$\frac{\text{Pelicans have } X.}{\text{Pelicans have } X.} \text{ is perfectly strong.}$$

According to the feature-based account, the strength of this argument is

$$\begin{aligned} a_x(\text{Pelicans/Pelicans}) &= \frac{\text{F(Pelicans)} \cdot \text{F(Pelicans)}}{|\text{F(Pelicans)}|^2} \\ &= \frac{\Sigma f_i(\text{Pelicans}) f_i(\text{Pelicans})}{|\text{F(Pelicans)}|^2} \\ &= \frac{|\text{F(Pelicans)}|^2}{|\text{F(Pelicans)}|^2} \\ &= 1. \end{aligned}$$

Certain unusual conditions exist which could cause a_x to take on values greater than 1. This will occur if the weight vector has the same or a similar direction as the input vector and a greater magnitude. An example of an argument leading to such a situation would be

$$\frac{\begin{array}{l}\text{Pelicans have } X. \\ \text{Flamingoes have } X.\end{array}}{\text{Pelicans have } X.}$$

The psychological strength of such an argument is not obvious, and may require detailed analysis. We can either ignore these degenerate cases or, as suggested earlier, extend the definition of a_x so that its maximum value is 1, with the result that no argument can be stronger than one with identical premises and conclusions.

PHENOMENA THAT DISTINGUISH THE TWO MODELS

The category-based and feature-based models differ with respect to three of the phenomena described by Osherson et al. (1990) as well as a new one.

xi. Premise–Conclusion Inclusion

The second limiting-case phenomenon stipulated by Osherson et al. (1990), and predicted by their model, states that arguments in which the conclusion category is subordinate to the premise category are perfectly strong:

$$\frac{\text{Animals have } X.}{\text{Mammals have } X.} \text{ is perfectly strong.} \qquad \text{(m)}$$

This phenomenon has an obvious logical justification. To the extent one knows and applies the category-inclusion rule, such arguments *are* perfectly strong. As phenomenon *vii,* the inclusion fallacy, suggests however, rules consistent with the logic of argument are not always applied. The feature-based account predicts that, to the extent that use of such rules is not overriding the postulated associative process, such arguments will not always be perfectly strong. Rather, they will depend on the featural overlap between premise and conclusion categories and the richness of the conclusion category representation. The strength of Argument m is

$$a_x(\text{Mammals/Animals}) = \frac{\text{F(Animals)} \cdot \text{F(Mammals)}}{|\text{F(Mammals)}|^2} \neq 1.$$

The numerator will increase as the extent of common features between the premise and conclusion categories increases, and therefore so will argument strength. The feature-based model predicts that, given most models of similarity, premise–conclusion inclusion arguments should vary with the similarity of premise and conclusion categories. The category-based model predicts that, since all such arguments are perfectly strong, no such variation should take place. A test of these different predictions is embodied in the experiments described in support of the next phenomenon.

Inclusion Similarity (New Phenomenon)

The strength of an argument in which the conclusion category is properly included in the premise category varies with the similarity of the two categories. I asked subjects to select the stronger of the following arguments:

1A.
Animals use norepinephrine as a neurotransmitter.

Mammals use norepinephrine as a neurotransmitter.

1B.
Animals use norepinephrine as a neurotransmitter.

Reptiles use norepinephrine as a neurotransmitter.

Forty-four of 50 subjects chose Argument A ($p < .001$). Also, Animals were judged significantly more similar to Mammals (mean similarity judgment was 5.7 on a 7-point scale) than to Reptiles (4.5 out of 7), $t(49) = 5.24$, $p < .001$. Subjects did not find these choices overwhelmingly difficult to make. The mean rating of confidence in choice of argument was 4.5 on a 7-point scale. I also asked subjects to choose between

2A.
Birds have an ulnar artery.

Robins have an ulnar artery.

2B.
Birds have an ulnar artery.

Penguins have an ulnar artery.

Again, the consistency in choice behavior was remarkable. Thirty-eight of 40 subjects chose A ($p < .001$). Also, Birds were judged significantly more similar to Robins (6.5) than to Penguins (4.6), $t(38) = 6.46$, $p < .001$. Again, subjects did not, on average, find the task extremely difficult. Mean confidence ratings in their choice were 4.6. Finally, and toward an examination of the phenomena in a different domain, I asked subjects which of the following they found stronger;

3A.
Furniture cannot be imported into Zimbabwe.

Tables cannot be imported into Zimbabwe.

3B.
Furniture cannot be imported into Zimbabwe.

Bookshelves cannot be imported into Zimbabwe.

Thirty-three of 39 subjects chose A ($p < .001$). Similarity judgments between Furniture and Tables (6.2) were significantly higher than those between Furniture and Bookshelves (5.3), $t(38) = 3.69$, $p < .001$. Again subjects were not overly strained by the choices they were asked to make; mean confidence was 3.9.

The preceding demonstrations suffer from a limitation imposed by the forced-choice task. The task required subjects to choose one or the other argument. Subjects could have based their choice on similarity, even though both arguments seemed perfectly strong, only to satisfy the task requirements. I therefore tried to replicate the inclusion similarity phenomenon using a rating task which did not make the task demands of the forced-choice procedure.

I gave 60 undergraduates at the University of Michigan the same six arguments as above (1A, 1B, 2A, 2B, 3A, and 3B) and asked them how convincing they found each one. To make their rating, subjects circled one of the integers from 1 (not at all convincing) to 10 (very convincing). In general, we would expect the strength of arguments satisfying premise–conclusion inclusion to be relatively high because categories usually share many features with their subordinates.

The mean convincingness rating for Argument 1A (Animals therefore Mammals) was 7.50, significantly greater than the mean rating for Argument 1B (Animals therefore Reptiles) of 5.88, $t(59) = 5.07$; $p < .001$. The same pattern held for the second pair of arguments (Birds therefore Robins versus Birds therefore Penguins). Mean convincingness ratings were 9.45 and 6.73, respectively, $t(59) = 7.35$; $p < .001$. The third pair also showed the phenomenon. The mean for 3A (Furniture therefore Tables) of 9.32 was significantly higher than the mean for 3B (Furniture therefore Bookshelves) of 8.53, $t(59) = 2.86$; $p < .01$. The argument with the more similar categories was judged significantly more convincing in all three cases. Similarity judgments replicated the patterns reported above.

Perhaps subjects failed to realize that each category was meant to subsume all members of that category. For example, they might have interpreted "Mammals have *X*" as "*some* Mammals have *X*." So I clarified the meaning of each statement by preceding each premise and conclusion category by the quantifier "all," for example "All animals use norepinephrine as a neurotransmitter." I asked a new group of 46 students to rate the convincingness of the six modified arguments.

I obtained an identical pattern of judgments. The mean convincingness ratings for the first two arguments (all Animals therefore all Mammals and all Animals therefore all Reptiles) slightly increased to 7.54 and 6.00, respectively, $t(45) = 3.02$; $p < .01$. For the second pair of arguments (all Birds therefore all Robins versus all Birds therefore all Penguins), corresponding means were 9.59 and 6.41, $t(45) = 6.76$; $p < .001$. Finally,

the mean for 3A (all Furniture therefore all Tables) was 8.35, whereas the mean for 3B (all Furniture therefore all Bookshelves) was 7.91, $t(45) = 3.24$; $p < .001$. Even when categories were explicitly quantified to include all category members, convincingness ratings were consistent with the model in all three cases.

The inclusion similarity phenomenon can be demonstrated in a different way. Instead of varying similarity by varying the conclusion category, we can vary the specificity of the premise category. Consider the two pairs of arguments

4A.
All birds require trace amounts of magnesium for reproduction.

All sparrows require trace amounts of magnesium for reproduction.

4B.
All animals require trace amounts of magnesium for reproduction.

All sparrows require trace amounts of magnesium for reproduction.

and

5A.
All dogs produce THS by their pituitary.

All German shepherds produce THS by their pituitary.

5B.
All mammals produce THS by their pituitary.

All German shepherds produce THS by their pituitary.

The first argument in each pair contains categories that share more common features than the corresponding second argument. This claim is supported by the similarity judgments of 44 subjects, who judged Birds and Sparrows (mean of 6.32) to be more similar than Animals and Sparrows (mean of 4.57), $t(43) = 7.76$; $p < .001$, and who also judged German shepherds to be more similar to Dogs (mean of 6.36) than to Mammals (mean of 4.59), $t(43) = 6.22$, $p < .001$. The feature-based model therefore predicts that subjects should rate Argument A of each pair as stronger. Indeed, the mean convincingness rating for 4A was 9.11, significantly greater than the mean for 4B of 8.11, $t(43) = 2.34$, $p < .05$. The premise specificity prediction was further supported by the second pair of arguments: 5A was rated significantly more convincing than 5B, mean ratings were 9.23 and 7.61, respectively, $t(43) = 3.17$, $p < .01$.

These demonstrations support the feature-based model which predicts the strength of arguments satisfying premise–conclusion inclusion to be proportional to coverage of the conclusion category's features by the premise category, and therefore correlated with similarity, and refute the

category-based model which cannot explain these data without auxilliary assumptions. One such auxilliary assumption[9] is to suppose that, in decomposing the premise category, subjects sample some of its subcategories and these tend to be the more typical. Upon incorporating this assumption, the category-based model will no longer predict premise–conclusion inclusion and will no longer always predict premise–conclusion identity.

xii and xiii. Nonmonotonicity (General and Specific)

As described with respect to the premise monotonicity phenomena (*v* and *vi*), the feature-based model never expects additional premises to decrease the strength of an argument. Osherson et al. (1990) show that this can happen however. Adding a premise that converts either a general or a specific argument to a mixed argument can decrease the strength of that argument. For example,

A.		B.
Crows have *X*.		Crows have *X*.
Peacocks have *X*.		Peacocks have *X*.
		Rabbits have *X*.
Birds have *X*.	is stronger than	Birds have *X*.

Similarly,

A.		B.
Flies have *X*.		Flies have *X*.
		Orangutans have *X*.
Bees have *X*.	is stronger than	Bees have *X*.

Osherson et al. (1990) explain these effects by invoking their concept of coverage. Flies cover the lowest-level category including Flies and Bees, namely Insects, better than Flies and Orangutans cover the lowest-level category including Flies, Orangutans, and Bees, namely Animals. A similar analysis applies to the other nonmonotonicity example. The feature-based model cannot explain the result.

One possible explanation for nonmonotonicities, consistent with a variant of the feature-based model, is that the features of the unrelated category compete with the features of the other premise categories. Features may be weighted by the number of premise categories that they are consistent with so that features shared by all categories would have the most influence on belief in the conclusion. Features appearing in only one

[9] Suggested by Ed Smith.

premise category of a multiple-premise argument could reduce argument strength if this category shared few features with (i) the other premise categories, for it would reduce the influence of their features, and (ii) the conclusion category, because it would then provide little additional feature coverage itself. Both of these conditions are met in the examples of nonmonotonicity above in that, for instance, Orangutans share few features with either Flies or Bees.

One implication of this hypothesis is that examples of nonmonotonicities should be observable in which the lowest-level inclusive category for both arguments is the same but the two feature-overlap conditions just described are met nevertheless. The similarity–coverage model could not explain such cases because it assumes that the lowest-level superordinate for the argument with the greater number of premises is at a higher level than the superordinate for the other argument. Preliminary data directed at this issue were collected by asking subjects to rate the convincingness of five pairs of arguments (see Appendix A). Each pair consisted of a single-premise specific argument such as

All crocodiles have acidic saliva.

All alligators have acidic saliva.

and a two-premise argument constructed by adding a premise category that, in the experimenters' judgment, had relatively few features in common with the other premise or the conclusion categories but came from the same lowest-level superordinate category as both the other categories (Reptiles in this example). For instance,

All crocodiles have acidic saliva.
All king snakes have acidic saliva.

All alligators have acidic saliva.

In each of the five cases nonmonotonicities were observed. The convincingness of the single-premise argument was rated as significantly higher (mean of 5.62) than that of the two-premise argument (mean of 4.98), $t(33) = 2.87$; $p < .01$. Nonmonotonicities were obtained using arguments whose categories apparently have the same lowest-level superordinate. To argue that subjects did use different (and nonobvious) superordinates for the arguments within each pair is to assume that category hierarchies are much less rigid and more idiosyncratic than the similarity–coverage model would suggest, at least in its simplest, most elegant form. If the hierarchies that subjects use are highly variable, and especially if they are context-dependent, then they may not provide any explanatory power beyond that of a featural description.

The feature-based model could be revised in a number of ways to

implement this "feature-competition" idea. One particularly simple way would be to allow for weight decay to be added to the encoding rule by introducing to Eq. 1 a forgetting parameter γ taking on some value between 0 and 1. The revised encoding rule would be

$$w_i(P_0,P) = \gamma w_i(P_0) + [1 - \gamma w_i(P_0)][1 - a_x(P/P_0)]a_i(P). \quad (11)$$

Equation 11 is a generalization of the original encoding rule in that we obtain Eq. 1 by setting $\gamma = 1$. It affords us a means of reducing the value of each weight by some proportion prior to its updating for each premise. Weight decay is useful when training feedforward connectionist networks for reasons other than those proposed here (Krogh & Hertz, 1992). The effect of weight decay on the feature-based model would be to reduce the influence of earlier premise features as subsequent ones are encoded. Weights corresponding to premise features unique to earlier premises would thereby decay while those common to earlier and later premises would have their representations bolstered, thus giving greater weight to features appearing in more than one premise.

The simplicity of the model of single-premise arguments, Eq. 5, is a direct result of the assumption that subjects know nothing about the relation between category features and the blank predicate, or that weights have initial value 0. According to Eq. 11, γ would thus be multiplied only by 0 when encoding the first premise. Therefore, adding weight decay would have no effect on the model of single-premise arguments. Careful choice of γ would lead to a model whose predictions were substantially unaffected in most other respects as well. For instance, we could make γ depend on the value of the output unit's current activation, while ensuring that its lowerbound was γ_0 and its upperbound was 1:

$$\gamma = \gamma_0 + \gamma_1 a_x(P/P_0),$$

where $0 < \gamma_0 < 1$ and $\gamma_0 + \gamma_1 = 1$. Such a model would have the following characteristics: (i) An immediately repeated premise would not affect argument strength. The strength of the argument P/C would be identical to that of P,P/C. (ii) We would expect nonmonotonicities whenever a premise category shared few features with both the earlier premise and the conclusion categories. For instance, given categories A, B, and C for which $F(A) \cdot F(B) \approx 0$ and $F(B) \cdot F(C) \approx 0$, the argument A/C would be stronger than the argument A,B/C. (iii) The only difference between this model and the original one would have to do with the weight each one gives to the different serial premise positions. This is discussed at greater length in the next section. The only phenomena that have been discussed that involve only multiple-premise arguments and are therefore necessarily affected by this change are diversity, feature exclusion, and of course

monotonicity and nonmonotonicity. Adding weight decay to the model would obviously decrease the relative frequency of monotonicities to nonmonotonicities. Accounts of diversity and feature exclusion can be developed for any order of premises, so changing the weight of the various serial positions would have unsystematic effects on these phenomena, effects that would be small in any case.

Order Effects

The feature-based and similarity-coverage models differ with respect to their predictions about the effect of premise order. Because the original feature-based model uses the delta rule of Eq. 1, premises are encoded only to the extent that they are surprising. Small differences would therefore sometimes be expected between an argument of the form A,B/C and one of the form B,A/C. For instance, if the set of features of the category in B were properly included in the set of features of the category in A, then the latter argument would be predicted to be slightly stronger than the former. The similarity–coverage model does not identify premises with their order of presentation, and so expects no effect of premise order.

However, adding weight decay to the feature-based model complicates its predictions concerning premise order. As well as expecting the surprisingness principle to sometimes decrease the impact of later premises, we now expect that weight decay will decrease the impact of earlier premises. Although predictions will depend to some degree on the extent of commonalities between all the categories within each argument, the major determinant of order effects will be the relative values of γ and the measure of surprisingness $(1 - a_x)$. More empirical work is necessary to sort these issues out.

QUANTITATIVE TESTS OF THE MODEL

To provide further empirical tests of the feature-based model, numerical predictions generated by the model were correlated with subjects' judgments of argument strength. Because the feature-based model has no free parameters, all we need to derive predictions is a set of categories, each with an associated feature list. Again, I used the ratings of strength of association between 85 properties and 48 mammals that Osherson et al. (1991) collected. These strength ratings can encourage subjects to judge a category as associated to properties that are not salient aspects of the subjects' conception of the category, as exemplified by the positive judgments between Sheep and *has tail* or *is weak*. Judgments of relative strength of association in this sense lead to some overestimation of the

value of features that are only weakly represented. I therefore applied a varying cutoff to the feature ratings, setting all values to 0 that were below a specified cutoff value.

Predictions of the feature-based model were calculated from the feature ratings using Eqs. 1, 2, and 3. Judgments of argument strength were obtained from data published in Osherson et al. (1990) and Smith et al. (1992). Five data sets were used, all involving arguments of the form described above. Categories were from the set of Mammals. The first set involved 15 two-premise-specific arguments, each with a different blank property (Smith et al., 1992, Table 4). The conclusion category was always Fox. Premise categories consisted of all possible pairs from the set (Dog, Hamster, Hippo, Mouse, Rhino, and Wolf). Because I did not have feature ratings for Dog, I used the ratings for German shepherd, on the assumption that a German shepherd is a highly typical dog. Smith et al. asked 30 University of Michigan undergraduates to estimate the probability of the conclusion on the assumption that the premises were true.

Cutoff values were varied from 0 to 1 in increments of 0.01. However, cutoff values close to 1 eliminated all or most of the data thereby rendering the feature ratings meaningless. I therefore did not consider cutoff values sufficiently close to 1 that category representations were all identical. Correlations were calculated between the model's predictions using feature ratings calculated for each cutoff value and subjects' probability estimates.

Feature ratings were obtained by averaging over subjects at one university (MIT) and argument strength ratings by averaging over subjects at a different university (Michigan). The only parameter varied was the cutoff value. Nevertheless, the mean correlation was 0.91 between the model's predictions and mean probability judgments, taken over all cutoff values less than 0.71. The maximum correlation was 0.96, achieved at cutoffs of 0.58 and 0.59. The magnitude of these correlations bodes well for the feature-based model.

By way of comparison to the similarity-coverage model, Smith et al. (1992) report a multiple correlation of 0.93 between (i) the same strength judgments as above and (ii) similarity and coverage terms estimated from a different group of 30 University of Michigan undergraduates. Strictly speaking, the correlations obtained by the two models are incomparable. The similarity–coverage model derives argument strength predictions from similarity judgments. Both argument strength and similarity ratings require subjects to directly compare premise and conclusion categories. The feature-based model derives argument strength predictions from feature ratings. Unlike argument strength ratings, feature ratings require subjects to evaluate the constituent attributes or components of a category. So judging argument strength and similarity are much more similar

tasks than are judging argument strength and rating features. Because correlations are partly just reflections of the similarity between two tasks, we would expect those obtained for the similarity–coverage model to be higher. In Osherson et al. (1991), similarities were derived from the same feature vectors as those I employ. The maximum correlation reported between theoretical and empirical similarity judgments was 0.64. Using these theoretical (feature-based) similarity judgments to predict argument strength judgments would surely have appreciably reduced the correlations obtained by the similarity–coverage model.

The second data set examined was identical to the first (probability judgments collected from the same 30 subjects), except that the conclusion category was Elephant instead of Fox (Smith et al., 1992, Table 4). The mean correlation between the data and predictions of the feature-based model was 0.86, taken over all cutoffs below 0.78. The maximum correlation was 0.97, achieved at cutoffs around 0.76. Smith et al. (1992) report a multiple correlation of 0.96 between the two estimated terms of their model and probability judgments.

The third data set again consisted of two-premise-specific arguments (Osherson et al., 1990, Table 5). This time the conclusion category was Horse. Only those arguments were considered that used premise categories for which I had feature ratings (Chimp, Gorilla, Mouse, Squirrel, Seal, Elephant, and Rhino). This allowed me to model 21 of 36 arguments reported by Osherson et al. (1990). Argument strength data consisted of mean rankings from 20 subjects of 36 arguments in terms of the likelihood of the conclusion given the premises.

Correlations between the model's predictions and mean argument rankings were substantially lower than the correlations reported above in which data came from probability judgments, but still significantly positive. The mean correlations for all cutoff points below 0.73 was 0.29. The maximum was 0.59 (at cutoff 0.60). Osherson et al. (1990) were able to calculate the multiple correlation between their estimated similarity and coverage terms and confirmation scores based on all 36 arguments. Similarity and coverage terms were estimated from similarity rankings provided by a different group of 40 subjects. They obtained a multiple correlation of 0.96, substantially higher than those found with the feature-based model. Their correlation could be higher because their data were more reliable due to the larger number of arguments they modeled.

Two data sets were available for quantitative tests of the model for general arguments. Osherson et al. (1990, Table 4) report confirmation scores for 45 three-premise general arguments in which the conclusion category was Mammal. Scores were mean rankings from 20 subjects over all 45 arguments of the likelihood of the conclusion given the premises. I had feature ratings for premises from 21 of these arguments. Feature

values for the category Mammal were estimated by assigning feature i of category Mammal the maximum value of feature i over the available set of 48 Mammals.[10]

Based on 21 arguments, the mean correlation was 0.71, taken over cutoffs less than 0.73. The maximum correlation was 0.83, which occurred at cutoff 0.29. Based on all 45 arguments, Osherson et al. obtained a correlation between their model's predictions and the data of 0.87, using the same similarity scores as those of the immediately preceding experiment.

Finally, probability estimates for 15 two-premise general arguments in which the conclusion category was Mammal were reported by Smith et al. (1992, Table 2). Estimates were made by 30 University of Michigan undergraduates. Again, I substituted German shepherd for Dog and used the maximum rule to obtain feature values for Mammal.

The mean correlation was 0.56 across all cutoff scores below 0.78. The maximum correlation obtained was 0.77, occurring at cutoff 0.04. The correlation obtained by Smith et al. between their model's predictions estimated from similarity ratings provided by a different group of University of Michigan students and probability estimates was higher, 0.92.

The correlations between model and data lend credence to the feature-based model for several reasons, despite the similarity–coverage model's higher showing in 3 out of 5 cases. First, as noted above, the similarity–coverage model derives its predictions from similarity judgments of objects at the same ontological level as argument strength judgments, namely categories, and the feature-based model derives them from objects at a reduced descriptive level, namely features. This gives the similarity–coverage model a relative advantage. Second, the data were not used to estimate any parameter intrinsic to the model, neither a learning nor an activation parameter. One parameter was used, one that affected which of 85 features were used to represent categories. To model specific arguments, the similarity–coverage model requires a parameter specifying the relative weight of the similarity and coverage terms. Third, feature values which provided the feature-based model's predictions were estimated using subjects from one population, whereas ratings of argument strength were provided by subjects from a different population. Fourth, for reasons given above, the feature ratings were not optimal for testing the feature-based model. Fifth, lacking subjects' direct judgments, feature ratings had to be estimated for two categories (Dog and Mammal). Fi-

[10] The use of mean feature ratings for Mammal had little impact on the correlations obtained. However, because of the large number of feature values of 0, mean ratings tended to give Mammal very small feature values which caused argument strengths to be too high, rarely less than 1.

nally, correlations for the feature-based model were generally high across a spectrum of cutoff values, as attested to by the generally high means.

DISCUSSION

I have described a simple model that accounts for a variety of phenomena regarding people's willingness to affirm a property of some category given confirmation of the property in one or more categories. The model implements the principle that one's willingness is given by the proportion of the category's features that are associated with the property; i.e., the proportion of features that are covered by the categories known to possess the property. The model accounts for 10 phenomena described by Osherson et al. (1990). It motivated empirical work suggesting that another one of their phenomena, premise–conclusion inclusion, does not hold in general although a related phenomenon predicted by the current model only, inclusion similarity, does. An experiment was reported to suggest that a generalization of the feature-based model may provide a more direct account of the two remaining phenomena, the nonmonotonicities. The model successfully predicted two new phenomena: feature exclusion, a reversal of the diversity phenomenon in which one of the more diverse premises fails to contribute to feature coverage, and category richness, in which an argument with a lower magnitude conclusion category is judged stronger than one with a higher magnitude. Preliminary data were also reported that were in a direction supporting the feature-based model's explanation for the asymmetry phenomenon. Finally, the feature-based model showed high correlations between its predictions and ratings of argument strength.

The feature-based model has two primary components that are probably not equally important in determining argument strength. Feature overlap, the dot product term in the feature-based model, seems to be generally more influential than representational richness, the magnitude term. Feature overlap played a role in all but one of the demonstrations of the various phenomena, premise–conclusion asymmetry. On the other hand, only four of the phenomena depended on vector magnitude, namely, category richness, typicality, asymmetry, and premise–conclusion identity. The relative weight of the dot product and magnitude terms in determining argument strength is related to the ratio of their variabilities. If, for instance, the magnitude of every representation were the same, then magnitude would play no role in determining argument strength. The variability in feature overlap among arguments may well be substantially greater than the variability in representational richness. The range of magnitude variation could be restricted by limited attention which may act to constrain the number of salient features of a category.

The Feature-Based View Versus the Category-Based View

Broadly construed, the category-based view of induction is not necessarily inconsistent with the feature-based view. Even if a feature-based model were consistent with a wide variety of data under many different conditions, a category-based model might still capture, in an easy-to-understand way, some large and important set of phenomena. Categories surely provide a useful level of abstraction. The category-based view has the distinct advantage of requiring only pairwise similarity ratings to generate quantitative predictions. The feature-based view is limited by the difficulties inherent in gathering reliable and complete sets of feature values for individual categories. On the other hand, the category-based view, by virtue of its abstractness, may be unable to capture all the subtleties of induction. A category-based model which appeals to relations among categories is not necessarily derivable from a feature-based model which appeals to relations among category attributes.

In terms of the specific models of concern here, the category-based and feature-based models differ in several respects. First, they suggest different ways of classifying the phenomena. The similarity–coverage model motivates a distinction between phenomena whose explanation relies on similarity—the prototype being the similarity phenomenon—versus phenomena whose explanation is premised on differences in coverage—the prototype being diversity. The feature-based model attributes both of these phenomena to a single factor, the degree of match between the premise categories as encoded in the weights and the conclusion category. A distinction more compatible with the feature-based model would separate phenomena explained in terms of the relative degree of match between weights and conclusion categories, the term in the numerator of Equation 3, with phenomena explained in terms of the relative degree of richness of the conclusion category, the denominator of the argument strength model. The asymmetry phenomenon is the prototype of the latter class.

Second, the two models differ with respect to their toleration of flexibility in our representations of categories. The feature-based model allows for extreme flexibility; the set of features representing a category could be highly context-dependent. By assuming a stable category hierarchy, the category-based model expects some rigidity. To the extent that new categories must be identified to explain new data, the usefulness of the similarity–coverage model is suspect. For example, what is the lowest-level category that includes ducks, geese, and swans? Is it birds, water-birds, web-footed birds? How do we deal with categories that are not natural kinds such as ad hoc categories (Barsalou, 1985), like birds found in national parks? The feature-based model requires only a set of

features for each category. On the other hand, to the extent we can appeal to an established category hierarchy, the featural description of individual categories might become unnecessarily tedious.

The models also differ with respect to the explanatory burden they put on different parts of an argument. Both models emphasize the relation between premise and conclusion categories. However, whereas the feature-based model puts some weight on characteristics of the conclusion category in isolation (by computing its magnitude), the similarity–coverage model gives weight to characteristics of the premise categories (by computing their coverage of a category that includes them).

Finally, the feature-based model has a computational advantage. All the model needs in order to derive argument strengths for a new category are its features. The similarity–coverage model requires similarity ratings (possibly derived from features) between the new category and each old one at the same hierarchical level, and the re-calculation of coverage values for all categories that include the new one.

In sum, we may be mistaken to interpret the success of Osherson et al.'s (1990) similarity–coverage model as implying the existence of two psychological processes which operate on pairwise similarity judgments, one of which computes overall similarity and the other coverage. Their pair of theoretical constructs may be only approximate abstract descriptions of more microlevel processing mechanisms.

The Problem of Nonblank Predicates

Some nonblank predicates do not seem fundamentally different from blank ones. Sloman and Wisniewski (1992) show that familiar predicates behave like blank ones when subjects cannot explain their relation to the categories of an argument. But cases of this sort are probably rare. The most important advantage of the feature-based view, and of the model proposed here in particular, is the direction it suggests for generalizing to arguments involving a range of nonblank predicates. One such direction would be to model some nonblank predicates as consisting of features, some of which have preexisting connections to features of premise and conclusion categories. One implication of this move would be that premises would influence argument strength only to the extent that they are surprising in relation to prior knowledge; that they inform people of links between features of conclusion categories and predicates that they do not already know.

A second implication of this type of feature-based model for nonblank predicates is that arguments will tend to be judged strong given a strong prior belief in their conclusion (for supporting evidence, see Lord, Ross, & Lepper, 1979). The model expects argument strength to be high whenever the weight vector is highly correlated with the conclusion category

vector, whether those weights come from the premises of the argument or from prior knowledge. The model predicts, therefore, that people will have trouble ignoring prior knowledge. (This is irrelevant when predicates are blank because, by definition, no prior beliefs exist.) However, practiced subjects may be able to evaluate a conclusion before and after encoding premises, and use the difference between the two outcomes as their measure of argument strength. Moreover, arguments like

Tables have legs.

Therefore, people have legs.

may be judged weak despite prior belief in the conclusion because the meager amount of overlap between the features of tables and those of people is so obvious.

A second direction involves generalizing the notion of feature coverage. Some nonblank predicates may have the effect of selecting those features of a category that are particularly relevant. Consider the argument

Tolstoy novels make good paperweights.

Michener novels make good paperweights.

Whatever your feelings and knowledge about literature, the only features of both kinds of novels that are relevant to this argument have to do with size, shape, and weight. An important and unanswered question is how those features become available to us so effortlessly. Indeed, some feature selection of this sort may be going on even with the "blank" predicates of this paper. The predicates are just about all biological in kind, not really entirely blank after all, and so other biological properties of the categories may have greater weight than nonbiological properties. This may be why we seem, sometimes, to respond to these arguments on the basis of taxonomic knowledge. Animal taxonomies are certainly informative about biological properties. But once the relevant features for each category of an argument are selected, feature coverage may become the principle determining argument strength. (The example argument is a strong one because the relevant features of Michener novels are covered by Tolstoy novels.) And the feature-based model would still be a contender as a way to implement that principle. These issues will have to be clarified before we can expect a feature-based model for nonblank properties to offer much insight.

CONCLUSION

The feature-based model provides for a new perspective on argument strength. By using rules normally associated with models of learning (the

delta rule and activation rule), it gives substance to the view that the process of confirming an argument is intimately related to concept learning. Encoding a premise is viewed as a process of linking new properties with old concepts. Testing a conclusion is described as a process of examining the extent to which the new links transfer to other concepts. The study of confirmation, for better or for worse, becomes an aspect of the study of generalization.

APPENDIX A

Thirty-four undergraduates from two psychology classes at the University of Michigan participated for prize money (a lottery was held after each session). They were tested in two groups of 17 students each. Each subject rated the strength or convincingness of 57 arguments. Twenty arguments were constructed to examine the effect of conclusion magnitude, 12 to investigate asymmetries, and 10 for nonmonotonicities. The remaining 15 arguments were used to investigate phenomena not reported in this paper. Every category of every argument was preceded by the word "all." Each premise was preceded by "FACT:" and each conclusion by "THEREFORE:"

Subjects received the following instructions: "Each piece of paper you have has an argument on it; i.e., a fact or series of facts followed by a conclusion. Please rate the convincingness of each of the arguments on a scale from 0 to 10. A rating of 0 means you find the argument completely unconvincing (or very weak) whereas a rating of 10 means you find the argument completely convincing. Do not be concerned if some of the terms seem unfamiliar to you, just rate the strength of each argument to the best of your ability. Each argument should be treated separately. In each case, please assume that the facts you are given are true. We would like to know how much those facts lead you to believe the conclusion." Special care was taken to separate the arguments within each pair used to test asymmetry and nonmonotonicity because the structure of these pairs is necessarily more apparent to subjects than the structure of the conclusion magnitude pairs. The asymmetry and nonmonotonicity arguments were combined with 11 other arguments and divided into two groups (groups A and B), with half of the arguments for each phenomenon in each group. The order of arguments within each of these groups was counterbalanced such that each argument appeared approximately an equal number of times in each serial position. The order in which phenomena were tested was the following: Subjects saw 10 conclusion magnitude arguments, 2 fillers, group A, 10 more conclusion magnitude arguments, 2 more fillers, and then Group B.

APPENDIX B

I derive the model of argument strength for two premise arguments involving categories D, E, and C,

$$\frac{\begin{array}{l}\text{D have } X.\\ \text{E have } X.\end{array}}{\text{C have } X.}$$

Equation 4 tells us that, after encoding premise category D, the weights are $W(D) = F(D)$. The activation of unit X upon presentation of category E is, from Eq. 5,

$$a_x(E/D) = \frac{F(D) \cdot F(E)}{|F(E)|^2}.$$

Using Eqs. 1 and 2, each weight i after having encoded both premise categories is

$$\begin{aligned} w_i(D,E) &= w_i(D) + [1 - w_i(D)][1 - a_x(E/D)]\, f_i(E) \\ &= f_i(D) + [1 - f_i(D)][1 - a_x(E/D)]\, f_i(E). \end{aligned}$$

To calculate the strength of the argument, we see how much category C now activates unit X:

$$\begin{aligned} a_x(C/D,E) &= W(D,E) \cdot F(C)/|F(C)|^2 \\ &= \frac{\Sigma f_i(D) f_i(C) + \Sigma \{[1 - f_i(D)][1 - a_x(E/D)] f_i(E) f_i(C)\}}{|F(C)|^2} \\ &= \frac{F(D) \cdot F(C) + [1 - a_x(E/D)][F(E) \cdot F(C) - \Sigma f_i(D) f_i(E) f_i(C)]}{|F(C)|^2}. \end{aligned}$$

REFERENCES

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

Gelman, S. (1988). The development of induction within natural kind and artefact categories. *Cognitive Psychology*, **20**, 65–95.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, **117**, 227–247.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart, J. L. McClelland, & The PDP Research Group (Eds.) *Parallel distributed processing* (Vol. 1). Cambridge, MA: The MIT Press.

Kahneman, D., & Tversky, A. (1973). The psychology of prediction. *Psychological Review*, **80**, 237–251.

Krogh, A., & Hertz, J. A. (1992). A simple weight decay can improve generalization. In

J. E. Moody, S. J. Hanson, & R. P. Lippmann, (Eds.) *Advances in neural information processing systems* (Vol. 4). San Mateo, CA: Morgan Kauffmann.

Lord, C., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, **37**, 2098–2109.

Osherson, D., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, **97**, 185–200.

Osherson, D., Stern, J., Wilkie, O., Stob, M., & Smith, E. E. (1991). Default probability. *Cognitive Science*, **15**, 251–269.

Quattrone, G. A., & Jones, E. E. (1980). The perception of variability within in-groups and out-groups: Implications for the Law of Small Numbers. *Journal of Personality and Social Psychology*, **38**, 141–152.

Rips, L. (1990). Reasoning. *Annual Review of Psychology*, **41**, 321–353.

Rips, L. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior*, **14**, 665–681.

Rothbart, M., & Lewis, S. (1988). Inferring category attributes from exemplar attributes: Geometric shapes and social categories. *Journal of Personality and Social Psychology*, **55**, 861–872.

Shafir, E., Smith, E. E., & Osherson, D. (1990). Typicality and reasoning fallacies. *Memory & Cognition*, **18**, 229–239.

Sloman, S. A., & Wisniewski, E. (1992). Extending the domain of a feature-based model of property induction. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Erlbaum.

Smith, E. E., Lopez, A., & Osherson, D. (1992). Category membership, similarity, and naive induction. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.) *From learning processes to cognitive processes: Essays in honor of W. K. Estes* (Vol. 2), Hillsdale, NJ: Erlbaum.

Smith, E. E., Shoben, E. J., & Rips, L. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, **81**, 214–241.

Sutton, R. S., & Barto, A. G. (1981). Towards a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, **88**, 135–170.

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

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, **90**, 293–315.

(Accepted September 24, 1992)